Chapter – 1

Organization Profile

ORGANIZATION	**:**	Central Scientific Instruments Organization (CSIO)
DEPARTMENT	**:**	Agrionics
GUIDE	**:**	Babankumar S. Bansod, Senior Scientist

1.1 Council of Scientific & Industrial Research (CSIR):

Council of Scientific & Industrial Research (CSIR), India was constituted in 1942 as an autonomous body under the provision of the Registration of Societies Act XXI of 1860. After independence, the need for bettering the living standards of the common man by promoting industry and for helping the industry to solve its problems through stimulus of scientific research was greatly stressed. The Council, through its constituent laboratories, has helped the country in increasing the economic growth and industrialization.

The Council has also helped the creation of new schools of research and in enlarging facilities for research by means of grants, training of research personnel, etc. The main functions of the Council are:

- Promotion, guidance and coordination of scientific and industrial research in India including other institutions and financing the specific research activities.
- Scientific study of problems affecting industries and trade.
- Award of Research Fellowships.
- Utilization of the results of researches conducted under the Council towards the development of industries in India.
- The establishment, maintenance and management of laboratories, workshops and organizations to further scientific and industrial research.
- The collection and dissemination of information in regard not only to research but also to industrial matters generally.
- Publication of scientific papers.

- Other activities to promote generally the objects of resolution.

Council of Scientific & Industrial Research (CSIR), India is perhaps among the world's largest publicly funded R&D organization. Its chain of 38 world class R&D establishments with their 80 field stations spread across India are manned by 10,000 highly qualified scientists and engineers, besides 13,000 auxiliary and other staff. Its range of activities cover practically the entire spectrum of industrial R&D ranging from aerospace to mining to microelectronics to metallurgy and so on. CSIR is truly a global R&D resource as its patrons and partners hail from over 50 countries.

1.2. Central Scientific Instruments Organization (CSIO):

Central Scientific Instruments Organization (CSIO) is a premier national laboratory dedicated to research, design and development of scientific and industrial instruments. It is one of the constituent laboratories of the Council of Scientific & Industrial Research (CSIR), India .A multi-disciplinary and multi-dimensional apex industrial research & development organization the country. Established in October 1959, CSIO was chartered to stimulate the growth of indigenous instrument industry in the country through development of contemporary technologies and other scientific & technological assistance.

Fig 1: CSIO

Initially located at New Delhi, CSIO moved to Chandigarh in 1962– the City Beautiful in the north west of Delhi. CSIO Campus (spread over an area of approximately 120 acres) comprises of Office Buildings, R&D Laboratories, Indo-Swiss Training Centre and a Housing Complex. An austere four-storey building and the accompanying workshops were inaugurated in

December 1967. Another four-storey block was added in 1976 for housing R&D Divisions, Library, etc. During mid-eighties, the laboratory buildings and infrastructural facilities were modernized in order to gear the Institute towards taking up development projects in challenging and emerging areas of technology. A separate Administrative Block was inaugurated in September 1994.

With a view to meeting the growing demand of well trained instrument technologists, Indo-Swiss Training Centre (ISTC) was started in December 1963 with the co-operation of Swiss Foundation for Technical Assistance, Zurich, Switzerland. CSIO is a multi-disciplinary organization having well equipped laboratories manned by highly qualified and well trained staff with infrastructural facilities in the areas of microelectronics, optics, applied physics, electronics, mechanical engineering, etc.

Large number of instruments ranging from simple to highly sophisticated ones, have been designed and developed by the Institute and their no-how's have been passed on to the industry for commercial exploitation. Having contributed substantially towards the growth of the scientific instruments industry in the country, CSIO enjoys high degree of credibility among the users of the instruments as well as the instrument industry.

1.2.1. Mission:

To be a leader at national level for designing and developing scientific and industrial instrument systems and devices; play a lead role in providing repair , maintenance & calibration and training of instrument technologists and be a custodian of instrumentation activity in the country.

1.2.2. Main Areas of Activity:

- Research, design and development of scientific & industrial instruments, components and system
- Service, maintenance, testing and calibration of instruments / components
- Human resource development in the area of instrumentation
- Technical assistance to industry

1.2.3. Organization Structure:

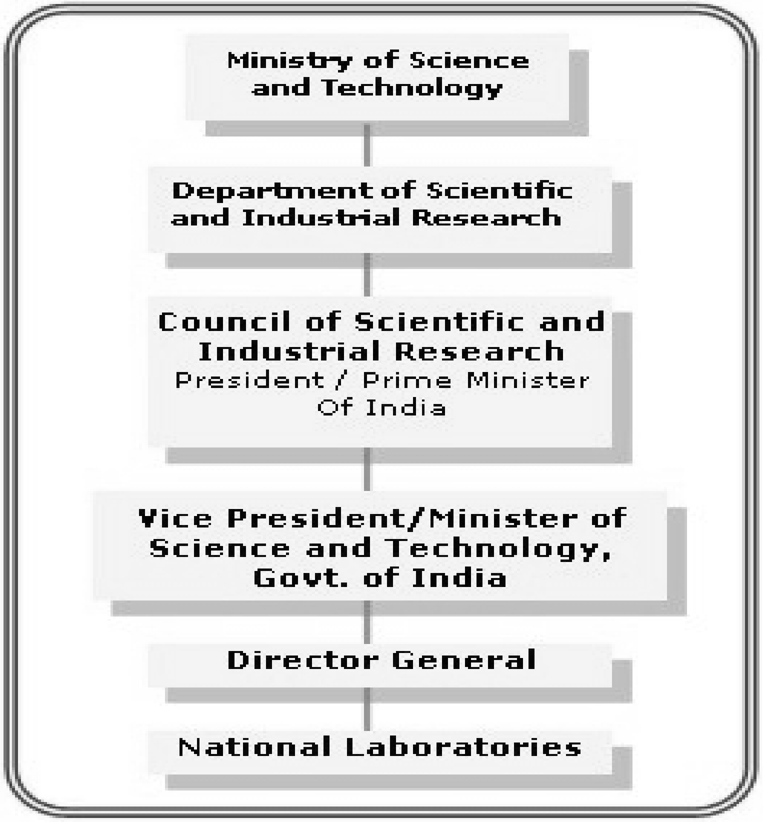

Ministry of Science and Technology
Department of Scientific and Industrial Research
Council of Scientific and Industrial Research
President / Prime Minister Of India
Vice President/Minister of Science and Technology, Govt. of India
Director General
National Laboratories

Fig 2: Position Flow Chart

1.2.4. Major R&D Areas:

- Strategic and defense applications
- Optics & opto-electronics
- Geo-scientific instrumentation
- Medical instrumentation
- Analytical instrumentation
- Agri-electronic instrumentation (Agrionics)
- Energy management, condition monitoring & quality control
- Environmental monitoring instrumentation
- Micro electro mechanical systems (mems) and sensors
- Biomolecular electronics and nanotechnology

1.3. CSIO Organizational Structure:

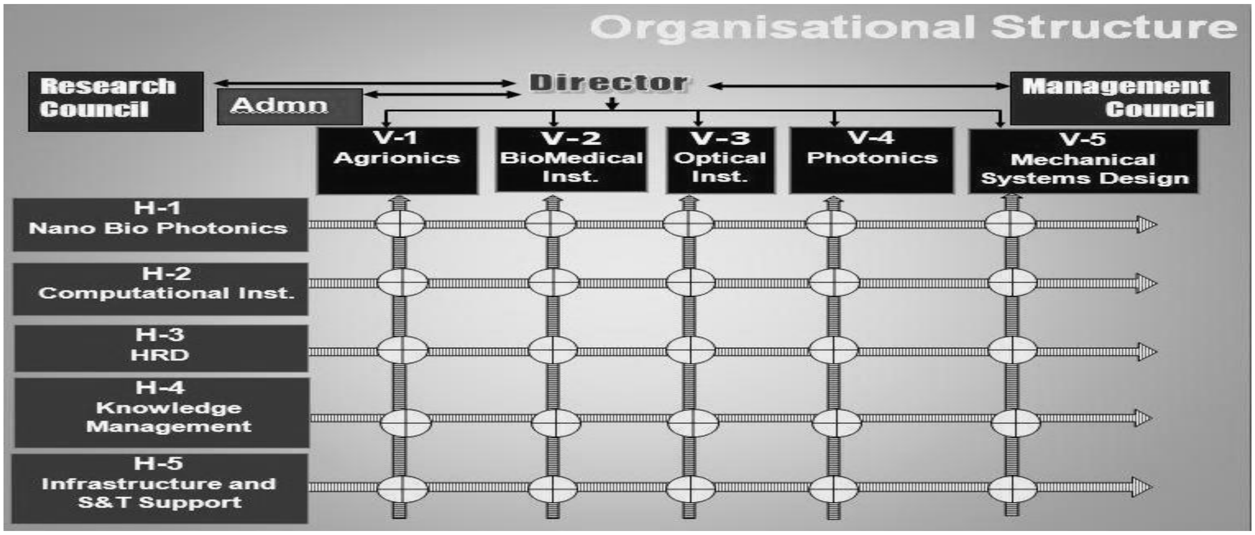


Fig. 3: CSIO Organizational Structure

1.4. Department Allotted (Agrionics):

The development of Agri-Electronic instruments involves a range of steps like a socialized man – power, a well equipped R&D center with manpower and latest modern equipment for calibration facilities in this area of instrumentation. The objective has been to take up the development of this technology and help in its production in India. The special outstanding feature of these developed technologies has been to provide complete range of field operable and portable instruments.

The areas identified for the instruments developed have been classified as: -

- Soil testing and nutrient measurement.
- Water and irrigation management.
- Instrumentation for horticulture and aquaculture.
- Crop rearing and food processing

Instruments developed:

Economic Field Usable pH Meter with Metallic Sensors, Inductive Electromagnetic Soil Salinity Tester , Digital Salinity Tester, Integrated System along with Specific Reagent s for the

Direct Measurement of Nitrite-n, Nitrate-N, Ammonia-N and Inorganic Phosphate-P, Specific Ion Analyzer, Portable Digital Soil Salinity Tester with 4-Electrode Probe, Iodine Value Meter for Edible Oils, Integrated System for Measurement of Metallic Impurities in Edible Oils (Oil Spectrophotometer), Digital Titrator Kit and Integrated System for the Measurement of Conductivity, pH mV & Temperature.

1.5. Technologies transferred to Industry:

1. Head-up display for LCA
2. Semi-Automatic Pick & Place Machine for Fine Pitch & Standard SMDs
3. Low Voltage Room Electrostatic Precipitator
4. Digital Indicator for Inclinometer/Tilt meter
5. Economic Field Usable PH Meter
6. Ophthalmoscope & Otoscope Diagnostic Set
7. Drug Infusion Pump & Controller
8. Pulse Oximeter
9. Resuscitation Bag (Ambu Bag) for Neonates
10. Servo Controlled Baby Care Incubator
11. Neonatal Oxygen Monitor
12. Single Puncture Laparoscope
13. Electronic Real Energy Meter (Three Phase and Single Phase)
14. Personal Dust Monitor
15. Nephelometer
16. On-line Analyzer for Energy Monitoring and Conservation
17. Portable Stack Opacity Monitor
18. Micro-Hardness Tester
19. Improved Lathe Tool Post
20. Glow Discharge Lamps

Chapter – 2

Introduction

2.1. Precision Farming:

Agricultural production system is an outcome of a complex interaction of seed, soil, water and agro-chemicals (including fertilizers). Therefore, judicious management of all the inputs is essential for the sustainability of such a complex system. The focus on enhancing the productivity during the green revolution coupled with total disregard of proper management of inputs and without considering the ecological impacts, has resulted into environmental degradation. The only alternative left to enhance productivity in a sustainable manner from the limited natural resources at the disposal, without any adverse consequences, is by maximizing the resource input use efficiency. It is also certain that even in developing countries, availability of labor for agricultural activities is going to be in short supply in future. The time has now arrived to exploit all the modern tools available by bringing information technology and agricultural science together for improved economic and environmentally sustainable crop production.

Precision agriculture merges the new technologies borne of the information age with a mature agricultural industry. It is an integrated crop management system that attempts to match the kind and amount of inputs with the actual crop needs for small areas within a farm field. This goal is not new, but new technologies now available allow the concept of precision agriculture to be realized in a practical production setting.

Precision farming is a term used to describe the management of variability within field boundaries, i.e. applying agronomic inputs in the right place, at the right time and in the right quantity to improve the economic efficiency and diminish the environmental impact of crop production [1]. Besides, it is in fact a comprehensive system designed to optimize agriculture production by carefully tailoring soil and crop management to fit the different conditions found in each field while maintaining environmental quality [2].

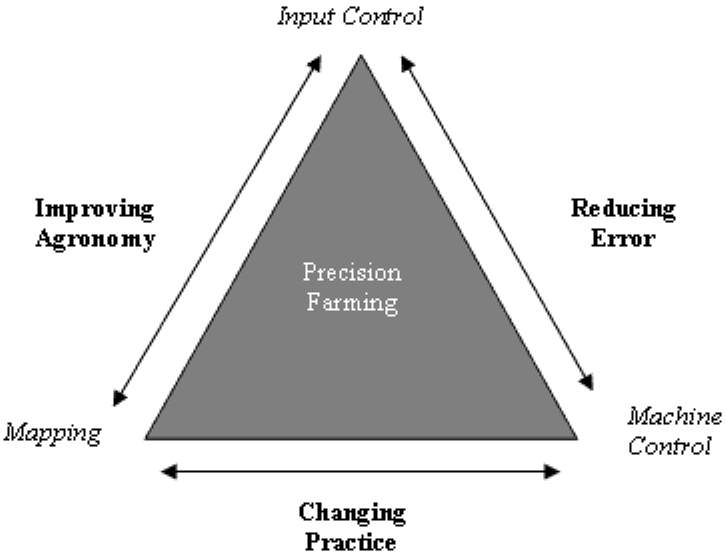


Fig. 4: Precision Farming Ideology

2.2. Definition of Precision Farming:

Precision Farming is generally defined as information and technology based farm management system to identify, analyze and manage variability within fields for optimum profitability, sustainability and protection of the land resource. In this mode of farming, new information technologies can be used to make better decisions about many aspects of crop production [3].

It involves looking at the increased efficiencies that can be realized by understanding and dealing with the natural variability found within a field. The goal is not to obtain the same yield everywhere, but rather to manage and distribute inputs on a site specific basis to maximize long term cost/benefit. Applying the same inputs across the entire field may no longer be the best choice. It is helping many farmers worldwide to maximize the effectiveness of crop inputs.

Precision agriculture often has been defined by the technologies that enable it and is often referred to as *GPS (Global Positioning System) agriculture or variable- rate farming*. As important as the devices are, it only takes a little reflection to realize that information earn higher returns than those who don’t. Precision farming distinguishes itself from traditional agriculture by its level of management wherein instead of managing whole fields as a single unit; management is customized for small areas within fields. This increased level of management emphasizes the need for sound agronomic practices. Before shifting to precision agriculture management, it is essential to have a good farm management system in place. Precision agriculture is a system approach to farming. To be viable, both economic and environmental benefits must be considered, as well as the practical questions of field level management and technologies needed (Figure 5). The issues related to precision agriculture include perceived benefits and also barriers to widespread adoption of precision agriculture management.

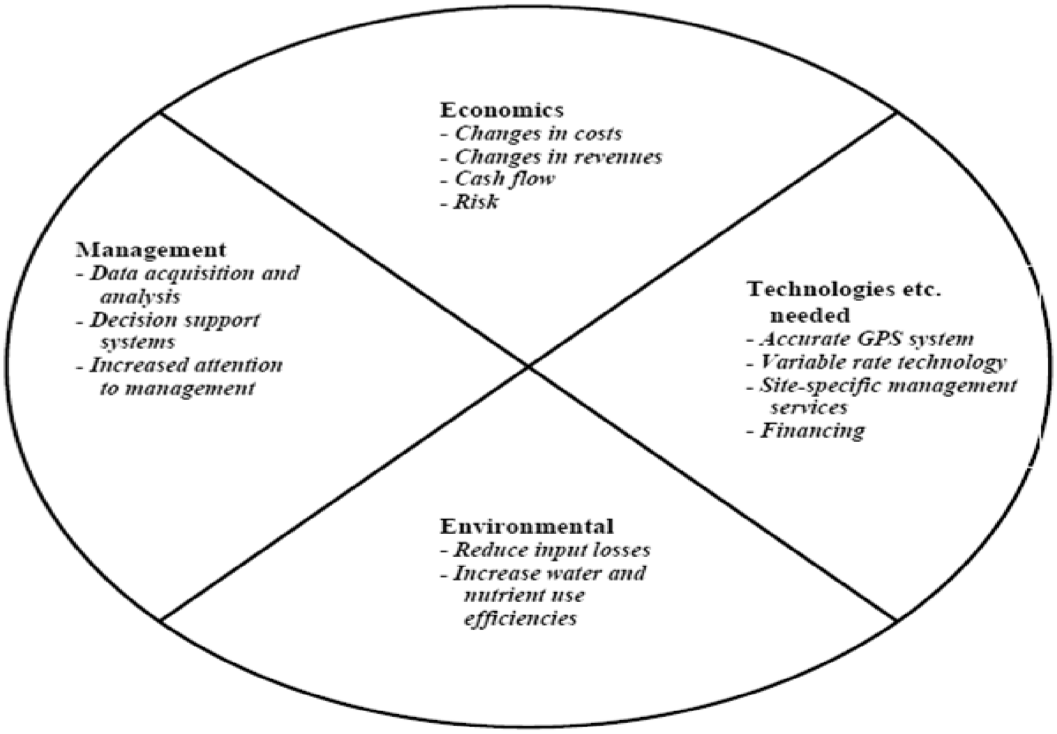

Economics
- Changes in costs
- Changes in revenues
- Cash flow
- Risk
Management
- Data acquisition and analysis
- Decision support systems
- Increased attention to management
Technologies etc. needed
- Accurate GPS system
- Variable rate technology
- Site-specific management services
- Financing
Environmental
- Reduce input losses
- Increase water and nutrient use efficiencies

Fig. 5: Important Issues related to Precision Farming

However, the conventional definition of precision farming is suitable when the land holdings are large and enough variability exists between the fields. In India, the average land holdings are very small even with large and progressive farmers. It is necessary to define revised definition of precision farming in the context of Indian farming while retaining the basic concept of precision farming.

The more suitable definition of precision farming in the context of Indian farming scenario could be: Precise application of agricultural inputs based on soil, weather and crop requirement to maximize sustainable productivity, quality and profitability. Today because of increasing input costs and decreasing commodity prices, the farmers are looking for new ways to increase efficiency and cut costs. Precision farming technology would be a viable alternative to improve profitability and productivity.

This does not mean having the same yield level in all areas of the field. A farmer's mental information database about how to treat different areas in a field required years of observation and implementation through trial and error. Today, that level of knowledge of field conditions is difficult to maintain because of the larger farm sizes and changes in areas farmed due to annual shifts in leasing arrangements. Precision agriculture offers the potential to automate and simplify the collection and analysis of information. It allows management decisions to be made and quickly implemented on small areas with larger fields.

2.3. Need For Precision Agriculture:

Precision agriculture gives farmers the ability to use crop inputs more effectively including fertilizers, pesticides, and tillage and irrigation water. More effective use of inputs means greater crop yield and/or quality, without polluting the environment. However, it has proven difficult to determine the cost benefits of precision agriculture management. At present, many of the technologies used are in their infancy, and pricing of equipment and services is hard to pin down. This can make our current economic statements about a particular technology dated. Precision agriculture can address both economic and environmental issues that surround production agriculture today. Questions remain about cost effectiveness and the most effective ways to use the technological tools we now have, but the concept of "doing the right thing in the right place at the right time" has a strong intuitive appeal. Ultimately, the success of precision agriculture depends largely on how well and how quickly the knowledge needed to guide the new technologies can be found.

What can precision farming do for me?

- Improve Crop Yield.
- Provide information to make better management decisions.
- Reduce chemical and fertilizer costs through more efficient application.
- Provide more accurate farm records.

- Increase profit margin.
- Reduce pollution.

2.4. Methods Used For Precision Farming:

Precision farming is a feasible approach for sustainable agriculture. Precision farming makes use of remote sensing to macro-control of GPS to locate precisely ground position and of GIS to store ground information. It precisely establishes various operations, such as the best tillage, application of fertilizer, sowing, irrigation, harvesting etc., and turns traditional extensive production to intensive production according to space variable data. Precision farming not only may utilize fully resources, reduce investment, decrease pollution of the of the environment and get the most of social and economic efficiency, but also makes farm products, the same as industry, become controllable, and be produced in standards and batches. However, precision farming has been confined to developed countries.

2.4.1. Two Methods of Precision Farming:

There are two methodologies for implementing precision, or site-specific, farming [4]. Each method has unique benefits and can even be used in a complementary, or combined, fashion:

1. **The first method:**

Map-based, includes the following steps: grid sampling a field, performing laboratory analyses of the soil samples, generating a site-specific map of the properties and finally using this map to control a variable-rate applicator (Figure. 6). During the sampling and application steps, a positioning system, usually DGPS (Differential Global Positioning System) is used to identify the current location in the field.

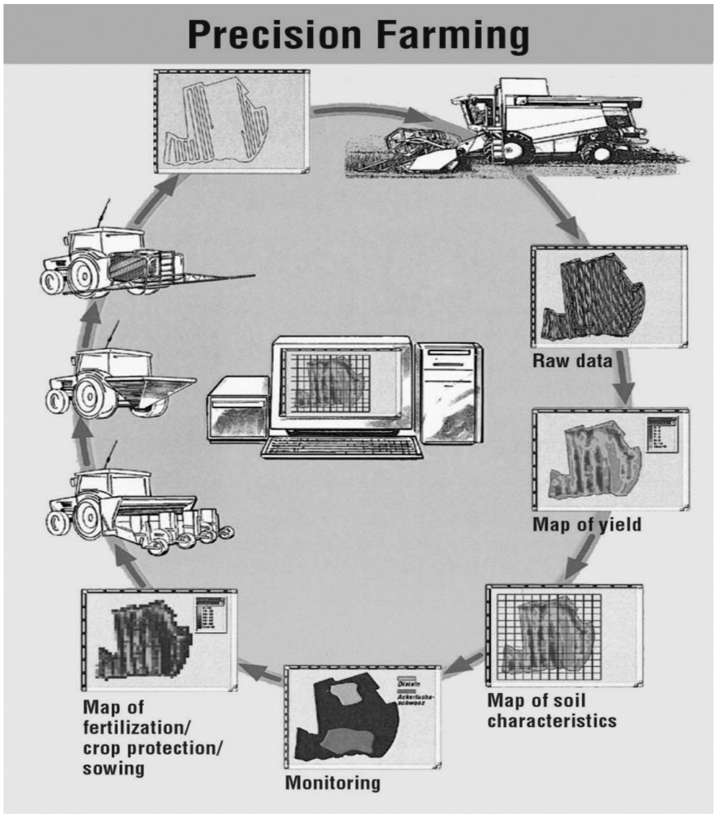

Precision Farming
Raw data
Map of yield
Map of soil characteristics
Monitoring
Map of fertilization/ crop protection/ sowing

Fig. 6: Map Based Method of Precision Farming

2. **The second method:**

Sensor-based, utilizes real-time sensors and feedback control to measure the desired properties on-the-go, usually soil properties, crop characteristics or soil nutrients, and immediately use this signal to control the variable-rate applicator. This second method doesn't necessarily require the use of a GPS system.

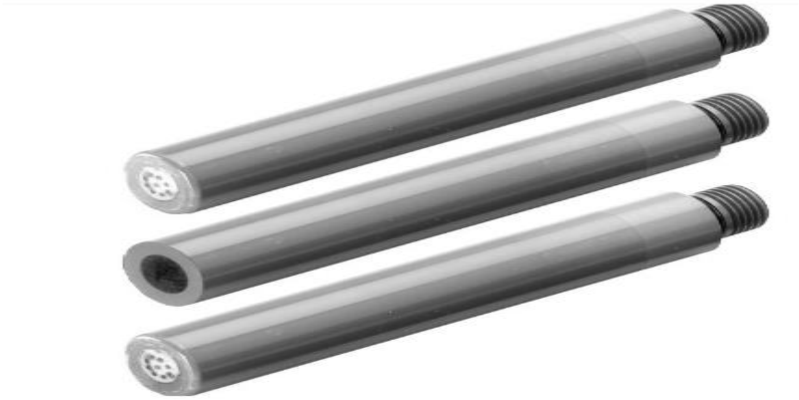


Fig. 7: Nutrient Ion Selective Electrode Sensors

2.5. Basic Steps in Precision Farming:

The basic steps in precision farming are [5]:

1. Assessing variation
2. Managing variation and
3. Evaluation

The available technologies enable us in understanding the variability and by giving site specific agronomic recommendations; we can manage the variability that make precision agriculture viable. And finally evaluation must be an integral part of any precision farming system. The detailed steps involved in each process are explained below:

- **Assessing:**

Assessing variability is the critical first step in precision farming. Since it is clear that one cannot manage what one does not know. Factors and the processes that regulate or control the crop performance in terms of yield vary in space and time. Quantifying the variability of these factors and processes and determining when and where different combinations are responsible for the spatial and temporal variation in crop yield is the challenge for precision agriculture. Techniques for assessing spatial variability are readily available and have been applied extensively in precision agriculture. The major part of precision agriculture lies in assessing to spatial variability. Techniques for assessing temporal variability also exist but the simultaneous reporting a spatial and temporal variation is rare.

- **Managing variability:**

Once variation is adequately assessed, farmers must match agronomic inputs to known conditions employing management recommendations. Those are site specific and use accurate applications control equipment. We can use the technology most effectively. In site-specific variability management, we can use GPS instrument, so that the site specificity is pronounced and management will be easy and economical. While taking the soil/plant samples, we have to note the sample site coordinates and further we can use the same for management. This results in effective use of inputs and avoids any wastage and this is what we are looking for.

For successful implementation, the concept of precision soil fertility management requires that within-field variability exists and is accurately identified and reliably interpreted, that variability influences crop yield, crop quality and for the environment. Therefore inputs can be applied accurately. The higher the spatial dependence of a manageable soil property, the higher the potential for precision management and the greater its potential value. The degree of difficulty, however, increases as the temporal component of spatial variability increases. Applying this hypothesis to soil fertility would support that Phosphorus and Potassium fertility are very conducive to precision management because temporal variability is low. For Nitrogen, the temporal component of variability can be larger than its spatial component, making precision Nitrogen management much more difficult in some cases.

- **Evaluation:**

There are three important issues regarding precision agriculture evaluation:

1. Economics
2. Environment and
3. Technology transfer

The most important fact regarding the analysis of profitability of precision agriculture is that the value comes from the application of the data and not from the use of the technology. Potential improvements in environmental quality are often cited as a reason for using precision agriculture. Reduced agrochemical use, higher nutrient use efficiencies, increased efficiency of managed inputs and increased production of soils from degradation are frequently cited as potential benefits to the environment.

2.6. Conclusion:

Precision Agriculture is the application of technologies and principles to manage spatial and temporal variability associated with all aspects of agricultural production for improving production and environmental quality. The success in precision agriculture depends on the accurate assessment of the variability, its management and evaluation in space-time continuum in crop production. The agronomic feasibility of precision agriculture has been intuitive, depending largely on the application of traditional arrangement recommendations at finer scales. The agronomic success of precision agriculture has been quite convincing in crops like sugar beet, sugarcane, tea and coffee. Successful implementation of precision agriculture depends on numerous factors, including the extent to which conditions within a field are known and manage, the adequacy of input recommendation and the degree of application control. The enabling technologies of precision agriculture can be grouped in to fine major categories: Computers, Global Positioning System (GPS), Geographic Information System (GIS), and Remote Sensing (RS) and Application control.

Precision farming basically depends on measurement and understanding of variability, the main components of precision farming system must address the variability. Precision Farming technology enabled, information based and decision focused, the components Include, (the enabling technologies) Remote Sensing (RS), Geographical Information System (GIS), Global Positioning System (GPS), Soil Testing, Yield Monitors and Variable Rate Technology.

Aspects of precision agriculture encompass a broad array of topics including variability of the soil resource base, weather, plant genetics, crop diversity, machinery performance and most physical, chemical and biological inputs used in crop production. Precision agriculture must fit the needs and capabilities of the farmer and must be profitable.

Chapter – 3

Work statement

The objective of the work titled *'Development of soil nutrient prediction model using pattern recognition techniques'* is to develop and test a real-time soil macronutrient analysis system, based on electrochemical impedance spectroscopy. The development of a real-time soil macronutrient sensor will allow the automated collection of soil nutrient data to accurately characterize within-field variability for site-specific fertilizer application.

The attempt is to develop a real time analysis system which will be able to predict accurately the level of macronutrients in the soil to a limited scale. Since there are many ions in the soil, we are focusing on three macronutrients i.e. Nitrate, Potassium & Phosphate.

The approach, electrochemical sensing based on ion-selective electrodes is used in real-time analysis because of its simplicity, portability, rapid response, and ability to directly measure the analyte with a wide range of sensitivity.

Impedance response of gold electrode for the samples in the frequency range of 1 Hz to 100 kHz has been measured. Frequency specific impedance response of these working electrodes along with their determined chemical concentrations was subjected to *'Principal Component Analysis'* to confirm the discriminability of the samples. The correlations between the frequency specific impedance response of working electrodes and the chemical concentrations which depend on sample variability have been established.

The samples have been discriminated and then are classified into clusters or classes using the classification technique *'Support Vector Machine'*. Each class has a distinct property that is similar for each member like concentration of the ions. Their responses are studied & patterns are identified.

The test samples are mapped to these classified classes using various regression techniques. The results are validated & are found to be correct to an approximate level. Cluster analysis is further conducted to interpret the results correctly.

The work is conducted successfully for lab testing, but the real soil samples needs to be tested which will confirm the accuracy & utility of this project. If the tests are conducted successfully, then the model will be programmed which will result in a real time sensor.

The application of the sensor will allow producers to identify fields and soils where variability can be profitably addressed by site-specific management technology. In addition, the sensor will provide accurate maps of soil nutrients, so that geo-referenced nutrient applications can be made with precision. Improved correlation between measured soil nutrient levels, nutrient removal due to harvested crops, and spatially applied nutrient additions will foster increased confidence in soil nutrient analysis. Agricultural producers will benefit from more efficient utilization of purchased inputs, and consumers will benefit from reduced adverse impact of agricultural practices on the environment.

Chapter – 4

Literature Review

4.1. Pattern Recognition:

Pattern recognition has undergone important developments for many years. Pattern recognition includes a lot of methods which are impelling the development of numerous applications in different fields. The practicability of these methods is intelligent emulation.

The design of a pattern recognition system requires careful attention to the following issues: definition of pattern classes, sensing environment, pattern representation, feature extraction and selection, cluster analysis, classifier design and learning, selection of training and test samples, and performance evaluation. [6]

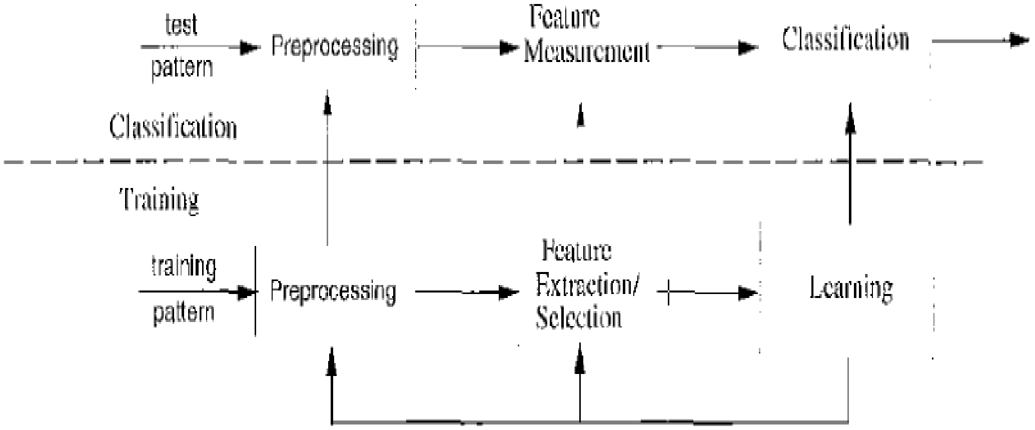


Fig. 8: Model for Pattern recognition

4.1.1. Definition of Pattern Recognition:

1978(Gonzalez, Thomas) defined Pattern recognition as a classification of input data via extraction important features from a lot of noisy data. [7]

2003(Sergios, Theodoridis) Pattern recognition is a scientific discipline whose aim is the classification of the objects into a lot of categories or classes. Pattern recognition is also a integral part in most machine intelligence system built for decision making. **[**8]

1992(Schalkoff) defined Pattern Recognition as “The science that concerns the description or classification (recognition) of measurements” [9]

4.1.2. Pattern Recognition Techniques:

- *Statistical pattern recognition*

Statistical decision and estimation theories have been commonly used in PR for a long time. It is a classical method of PR which was found out during a long developing process, it based on the feature vector distributing which getting from probability and statistical model. The statistical model is defined by a family of class-conditional probability density functions Pr(x|ci)(Probability of feature vector x given class ci) In detail, in SPR, we put the features in some optional order, and then we can regard the set of features as a feature vector. [10]Also statistical pattern recognition deals with features only without consider the relations between features.

- *Data clustering*

Its aim is to find out a few similar clusters in a mass of data which not need any information of the known clusters. It is an unsupervised method. In general, the method of data clustering can be partitioned two classes, one is hierarchical clustering, and the other is partition clustering.

- *Structural pattern recognition*

The concept of structural pattern recognition was put for the fourth time. [11] And structural pattern recognition is not based on a firm theory which relies on segmentation and features extraction. Structural pattern recognition emphases on the description of the structure, namely explain how some simple sub-patterns compose one pattern. There are two main methods in structural pattern recognition, syntax analysis and structure matching. The basis of syntax analysis is the theory of formal language, the basis of structure matching is some special technique of mathematics based on sub-patterns. When consider the relation among each part of the object, the structural pattern recognition is best. Different from other methods, structural pattern recognition handle with symbol information, and this method can be used in applications with higher level, such as image interpretation. Structural pattern recognition always associates with statistic classification or neural networks through which we can deal with more complex problem of pattern recognition, such as recognition of multidimensional objects.

- *Applications of Support Vector Machine (SVM) for pattern recognition*

SVM is a relative new thing with simple structure; it has been researched widely since it was proposed in the 1990's. SVM base on the statistical theory, and the method of SVM is an effective tool that can solve the problems of pattern recognition and function estimation, especially can solve classification and regression problem, has been applied to a wide range for pattern recognition such as face detection, verification and recognition, object detection and recognition ,speech recognition etc. [12]

4.1.3. Structure of pattern recognition system:

A pattern recognition system based on any PR method mainly includes three mutual-associate and differentiated processes. [13] One is data building; the other two are pattern analysis and pattern classification .Data building convert original information into vector which can be dealt with by computer. Pattern analysis' task is to process the data (vector), such as feature selection， feature extraction， data-dimension compress and so on. The aim of pattern classification is to utilize the information acquired from pattern analysis to discipline the computer in order to accomplish the classification.

A very common description of the pattern recognition system that includes five steps to accomplish. The step of classification/regression / description showed in fig 8 is the kernel of the system. Classification is a PR problem of assigning an object to a class. The output of the PR system is an integer label, such as classifying a product as "1" or "0" in a quality control test. Regression is a generalization of a classification task, and the output of the PR system is a real-valued number, such as predicting the share value of a firm based on past performance and stock market indicators. Description is the problem of representing an object in terms of a series of primitives, and the PR system produces a structural or linguistic description.

A general composition of a PR system is given below:

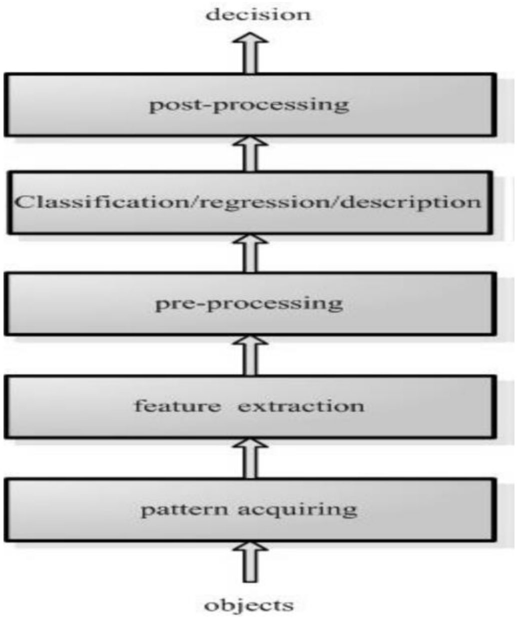


Fig. 8: Structure of Pattern Recognition

4.1.4. Classification Process in PR:

Any classification method uses a set of *features* or *parameters* to characterize each object, where these features should be relevant to the task at hand. We consider here methods for *supervised* classification, meaning that a human expert both has determined into what classes an object may be categorized and also has provided a set of sample objects with known classes.[14] This set of known objects is called the *training set* because it is used by the classification programs to learn how to classify objects. There are two phases to constructing a classifier. In the training phase, the training set is used to decide how the parameters ought to be weighted and combined in order to separate the various classes of objects. In the application phase, the weights determined in the training set are applied to a set of objects that do *not* have known classes in order to determine what their classes are likely to be.

A simple classification approach that is used is nearest neighbor classifier. In this, first a central median is calculated and based on the distance between the points and the medians, the cluster are formed. The different points are grouped into different clusters.

4.1.5. Data Mining & Clustering Techniques in PR:

Data mining is one of the areas where there is extensive use of pattern recognition. The two important tasks of Data mining are clustering and classification. [15] Clustering helps in finding patterns in the data that share commonalities and properties. There are different types of clustering used:

1. *K Means:*

In this the cluster number ‘k’ is pre specified and the data is grouped into those k clusters based on the distance between the mean and that point. Clustering based on k-means is closely related to a number of other clustering and location problems. These include the Euclidean k-medians in which the objective is to minimize the sum of distances to the nearest center and the geometric k-center problem in which the objective is to minimize the maximum distance from every point to its closest center. The user specifies the number of clusters in advance, and can also define cluster membership as well. The output is to give class identification for each object (sample). *K-Means methodology* is a commonly used clustering technique. The K-medians methodology is also used, and though slower than K-means, is more robust to outliers. In both cases the analysis involves starting with a collection of samples that one attempts to group them into *k* Number of clusters based on certain specific distance measurements.

2. *Conceptual clustering:*

Conceptual clustering is an advanced data mining technique that clusters data into clusters associated with conceptual representations, or conceptual clusters. Concept hierarchy can then be constructed from the conceptual clusters. However, traditional conceptual clustering techniques can only work on specific data types such as nominal and numeric. In addition, the concept hierarchy is mostly in a tree-like structure which is unable to support the representation of multiple inheritance.

Conceptual clustering techniques can be used to construct a concept hierarchy from data. Traditional conceptual clustering techniques such as COBWEB and AutoClass are based on taxonomy clustering techniques and use statistical models as conceptual representations of clusters. However, these techniques are only applicable to specific types of data.

3. *Fuzzy clustering:*

Fuzzy clustering methods, however, allow the objects to belong to several clusters simultaneously, with different degrees of membership. In many situations, fuzzy clustering is more natural than hard clustering. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. The discrete nature of the hard partitioning also causes difficulties with algorithms based on analytic functional, since these functional are not differentiable.

4.2. Principal component analysis:

PCA is a statistical tool commonly used for the unsupervised classification of multivariate data. [16] The main aim of principal component analysis is to reduce dimensionality of data, giving a small number of principle components (PCs) that represent the vast majority of variance in the data. PCA finds an alternative set of axes about which a data set may be represented and indicates along which axis there is most variation. This allows more effective visualization and classification of multivariate data. This method allows natural clustering of the data to be observed as well as providing an insight into how effectively a pattern recognition system could classify the data.

PCA is a statistical technique for the reduction of input data dimension and is largely used for feature extraction. [17] It captures the relevant information in a set of input data providing a lower dimension, but informative representation of the original data. It sequentially creates a set of principal components from the original data. The first principal component (PC1) maps the maximum variance and information of the input data followed by the other principal components (PC2, PC3 and so on) in descending order of the variance. Generally a good discrimination is mapped by the first two principal components, i.e. PC1 and PC2.

PCA is a dimension reduction method which uses correlated variables and identifies orthogonal linear recombination of the variables that summarize the principal sources of variability in the data. The input for analysis was a correlation matrix involving some selected variables, resulting in normalized PCA. The number of principal components is the same as that the number of considered variables, but usually the first few components explains most of the total variance in the data set. With the aim of generating homogeneous subfields, a fuzzy cluster analysis was performed, considering the factors, principal components, previously indicated. In this work, the FCM, an unsupervised continuous classification procedure, was used to divide the field into different cluster classes.[18] This classification algorithm is preferred for grouping properties in the soil continuum, because it produces a continuous grouping of objects by assigning partial class membership.

PCA is a statistical technique for the reduction of input data dimension and is largely used for feature extraction. [19] It captures the relevant information in a set of input data providing a lower dimension, but informative representation of the original data. It sequentially creates a set of principal components from the original data. The first principal component (PC1) maps the maximum variance and information of the input data followed by the other principal components (PC2, PC3 and so on) in descending order of the variance. Generally, a good discrimination is mapped by the first two principal components, i.e., PC1 and PC2.

4.3 Soil Macronutrient Testing:

The soil macronutrients, nitrogen (N), phosphorus (P), and potassium (K), are essential elements for crop growth. The application of commercial N, P, and K fertilizers has contributed to a tremendous increase in yields of agricultural crops that feed the world’s population. However, excessive use of these fertilizers has been cited as a source of contamination of surface and groundwater.[20,21] Ideally, application rates should be adjusted based on estimates of the requirements for optimum production at each location because there is high spatial variability of N, P, and K within individual agricultural fields.[22, 23]

Site-specific crop management (SSCM), also called precision agriculture, is a soil and crop management system that assesses variability in soil properties (e.g., pH, organic matter, and soil nutrient levels), field (e.g., slope and elevation) and crop parameters (e.g., yield and biomass), to optimize inputs such as fertilizers and herbicides based on information obtained at within-field locations.[24] SSCM aims to improve profitability and to better protect soil and water resources as compared to conventional management practices.[25]

Soil nutrient testing is a management tool that can help accurately determine the available nutrient status of soils and guide the efficient use of fertilizers. With the increasing awareness of fertilizer effects on environmental and soil quality, soil tests have been instrumental in determining where insufficient or excess nutrient levels occur. However, conventional soil testing methods, based on manual or mechanical soil sampling and colorimetric or atomic emission spectroscopy, are costly and time consuming. This expense limits the number of samples analyzed per field, making it difficult to characterize spatial or temporal variability in soil nutrient concentrations within fields. [26]

In particular, accurate monitoring of soil nitrate has been limited by the relatively long turn-around time of laboratory analysis, because soil nitrate can be easily lost by leaching and denitrification between the time of testing and plant uptake. Therefore, quantifying soil nitrate variability requires a fast on-site measurement at a high sampling intensity that will allow the variability to be mapped spatially and temporally with some degree of confidence. [27]

The time and cost required for the intensive sampling needed in SSCM, when using conventional sampling and analysis techniques, may make implementation of a variable-rate fertilizer application system impractical. In this situation, on-the-go real-time sensors could be useful to allow the collection of geographically referenced data on a much finer spatial resolution than is currently feasible with manual and/or laboratory methods. These automated sensor measurements can provide the benefits from the increased density of measurements at a relatively low cost. [28]

Most of the electrochemical methods used to determine soil nutrient levels are based on the use of an ion-selective electrode (ISE), with glass or a polymer membrane, or an ion-selective field effect transistor (ISFET). The ISFET has the same theoretical basis as the ISE, i.e., both ISEs and ISFETs respond selectively to a particular ion in solution according to a logarithmic relationship between the ionic activity and electric potential. [29] The ISEs and ISFETs require recognition elements, i.e., ion selective membranes, which are integrated with a reference electrode and enable the chemical response (ion concentration) to be converted into a signal (electric potential). Due to an increased demand for the measurement of new ions, and tremendous advances in the electronic technology required for producing multiple channel ISFETs, numerous ion-selective membranes have been developed in many areas of applied analytical chemistry, e.g., in the analysis of clinical or environmental samples. Ion-selective membranes are available for sensing most of the important soil nutrients, including NO3, K, Na, Ca, Mg, and Cl. [30-31]

On-site monitoring of N, P, and K nutrients is preferable due to the potential for a higher density of measurements at a relatively low cost, allowing more efficient mapping of soil nutrient variability for variable–rate nutrient management. Optical diffuse reflectance in visible and near-infrared wavelength ranges has been used as a non-destructive method to rapidly quantify soil properties for site-specific management. However, application of optical sensor technology for on-site measurements of soil nutrients has been limited, primarily due to relatively poor estimates at critical macronutrient levels for soil fertility management, as well as strong effects of soil type. In principle, electrochemical sensing with ion-selective electrodes or ion-selective field effect transistors is a promising approach for real-time analysis because of rapid response, direct measurement of the analyte with a wide range of sensitivity, simplicity and portability. The disadvantage of on-the-go sensors based on ion selective technology is that soil sampling and nutrient extraction are required, increasing the complexity of the system and the time required for a measurement. However, recent successful commercialization of a soil pH mapping system based on ion selective technology shows there is potential to overcome these issues. Improved on-the-go soil macronutrient sensing leading to potentially commercial products will require additional research and development efforts. First of all, in the near future, further efforts are needed to improve the durability of current on-the-go sensing systems under harsh conditions found in the field. The adoption of a simple alarm monitor to signal the operator in the event of a system malfunction may be the first step toward commercial success. For practical use, future systems that allow continuous monitoring of soil nutrients will require further research to integrate soil sample collection, automated sample preparation, and nutrient analysis. These systems will likely rely on technical advances in electronics and mechanical engineering such as microelectromechanical systems (MEMS) and flow injection analysis (FIA) that may enable the system to be fully automated, thereby providing high reproducibility in sensor data. Since these miniaturization techniques will accommodate low-volume samples, resulting in a reduction in reagent consumption and waste generation, a feasible automated soil sampler and extraction system may be easier to develop. Furthermore, use of a sensor array capable of determining several analytes, such as soil macronutrients and pH, simultaneously would further reduce sample processing time, sample volume and reagent consumption. Integration of the multiple data streams available from such a multianalyte sensor array might provide improved estimates of the individual analytes through its ability to quantify and factor out any cross-channel responses. Widespread adoption of on-the-go soil nutrient sensing may be somewhat limited by the degree to which precise sampling and rapid extraction of the macronutrients in the sample can be achieved in a real-time system. Because extraction efficiency is strongly affected by the extraction time and because the time required for complete extraction may not be feasible in a real-time system, this approach may provide different results as compared to traditional soil testing methods. In this regard, research will be needed to calibrate sensor-based nutrient measurements against plant nutrient response, so that agronomists and growers gain confidence in the applicability of the new methods. Such a calibration might be implemented in the same way that past calibrations to standard laboratory measurements were developed. However, this process would require numerous field experiments with different crops and soil types. An alternative method, whereby sensor measurements were directly calibrated to laboratory nutrient measurements across a broad range of conditions, might be preferable. Although the calibration to plant response would be an indirect one with this approach, it would be considerably less costly and time-consuming. [32]

4.3. Soil Sensors:

The basic objectives of site-specific management of agricultural inputs are to increase profitability of crop production, improve product quality, and protect the environment. Information about the variability of different soil attributes within a field is essential for the decision-making process [33]. The inability to obtain soil characteristics rapidly and inexpensively remains one of the biggest limitations of precision agriculture. Numerous researchers and manufacturers have attempted to develop on-the-go soil sensors to measure mechanical, physical and chemical soil properties. The sensors have been based on electrical and electromagnetic, optical and radiometric, mechanical, acoustic, pneumatic, and electrochemical measurement concepts.

4.3.1. Instrumentation & Methods:

The global positioning system (GPS) receivers, used to locate and navigate agricultural vehicles within a field, have become the most common sensor in precision agriculture. In addition to having the capability to determine geographic coordinates (latitude and longitude), high-accuracy GPS receivers allow measurement of altitude (elevation) and the data can be used to calculate slope, aspect and other parameters relevant to the landscape. When a GPS receiver and a data logger are used to record the position of each soil sample or measurement, a map can be generated and processed along with other layers of spatially variable information. This method is frequently called a “map-based” approach. [34]

- *Electrical and electromagnetic sensors*

Various measurement systems are based on electrical circuits and used to determine the ability of certain media to conduct or accumulate electrical charge. If soil is used as such a medium, its physical and chemical characteristics can affect circuit behavior and, thus, the measured electric parameters. Rapid response, low cost and high durability have made electrical and electromagnetic sensors the most attainable techniques for on-the-go soil mapping. Obtained maps have been correlated to: soil texture, salinity, organic matter, moisture content, and other soil attributes. [35]

- *Optical and radiometric sensors*

Measurement of reflectance, absorption or transmittance characteristics of a material provides a non-destructive and rapid technique to evaluate its properties. Determination of the amount of energy reflected from the soil surface in a particular spectral range is the most popular approach in agriculture. Similar to electrical and electromagnetic sensors, optical and radiometric measurements are frequently affected by a combination of soil attributes. However, the response in different parts of the spectral range may be affected by various soil properties to different degrees, which provides an opportunity to separate several effects with a single sensor response. Moisture, organic matter, particle size, iron oxides, mineral composition, soluble salts, parent material, and other attributes affect soil reflectance. [36]

- *Electrochemical sensors*

Most sensors described above have been used to directly or indirectly assess spatial variability of different mechanical and physical soil properties. On the other hand, direct on-the-go measurement of soil chemical characteristics, such as pH or nutrient content has been the objective of considerable research. Electrochemical methods have been success fully used to directly evaluate soil fertility. This is usually done by either an ion-selective electrode (glass or polymer membrane), or an ion-selective field effect transistor (ISFET). In both cases, measured voltage (potential difference) between sensing and reference parts of the system is related to the concentration of specific ions (H+, K+, NO3−, etc.). Ion selective electrodes have been historically used by commercial soil laboratories to conduct standard chemical soil tests, and they are widely used to measure soil pH. Ion-selective field effect transistors have several advantages over ion selective electrodes, such as small dimensions, low output impedance, high signal-to-noise ratio, fast response and the ability to integrate several sensors on a single electronic chip. [37]

Adsett et al. (1999) also developed a prototype soil nitrate monitoring system consisting of a soil sampler, a conveying and metering unit, an extraction and measurement unit, and a control unit. Through the analysis of response curves for different soil types, a procedure was developed to predict soil nitrate using an NO3− ion-selective electrode in less than 10 s. While obtaining acceptable results during laboratory testing, field evaluation revealed the need for additional improvement of the sampler and other system components. [38]

4.4. Clustering:

The aim of clustering analysis is to group a given data set into clusters. Aresulting partition should possess the following properties:

1. Homogeneity within the clusters, i.e., data that belong to the same cluster should be as similar as possible and
2. Heterogeneity between the clusters, i.e., data that belong to different clusters should be as different as possible [39].

Davis–Bouldin [40] presented a cluster separation measure which represents class separability by the ratio of the sum of intracluster scatter to intercluster separation. Since scatter matrices depend on the geometry of the clusters, this index has both a statistical and geometric rationale. Let be a cluster of and completely belongs to.

Then, the measure of scatter within can be obtained by (1) where the centroid of is and is the number of vectors in cluster. Similarly, the measure of separation between cluster and cluster , can be obtained by (2) where is the element of . The similarity between cluster and cluster can be obtained by (3) by assuming, we have adopted Euclidian distance as a measure between intracluster and intercluster separation.

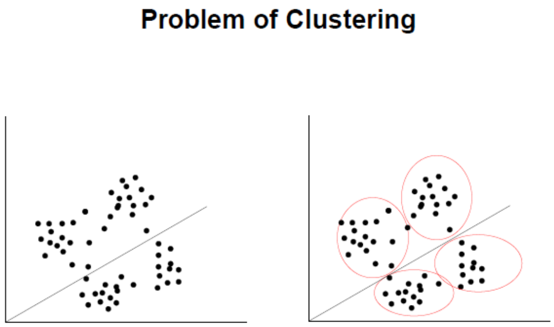


Fig. 9: Problem of clustering

4.4.1. Use of Clustering in Data Mining:

Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign. [41]

For example, a company that sale a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving fewer sales then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process.

4.5. Electrodes Used in Frequency Response Analyzer:

An electrode is a (semi-)conductive solid that interfaces with a (n) (electrolyte) solution. The common designations are:

- Working Electrode
- Reference Electrode and
- Counter Electrode

Working electrode is the designation for the electrode being studied. In corrosion experiments, this is likely the material that is corroding. In physical chem. experiments, this is most often an inert material— commonly gold, platinum or carbon—which will pass current to other species without being affected by that current.

The Counter or Auxiliary electrode is the electrode in the cell that completes the current path. All electrochemistry experiments (with non-zero current) must have a working – counter pair. In most experiments the Counter is simply the current source/sink and so relatively inert materials like graphite or platinum are ideal, though not necessary. In some experiments the counter electrode is part of the study and the material composition and setup will vary accordingly.

Reference electrodes are, as their name suggests, electrodes that serve as experimental reference points. Specifically they are a reference for the potential (sense) measurements. Reference electrodes should, therefore, hold a constant potential during testing, ideally one which is known on an absolute scale. This is accomplished by first having little or, ideally, no current flow through them, and second by being “well poised” which means that even if some current does flow it will not affect the potential. While many electrodes could be well poised there are several that are very commonly used and commercially available: Silver/Silver Chloride, Saturated Calomel, Mercury/Mercury (mercurous) Oxide, Mercury/Mercury Sulfate, Copper/Copper Sulfate, and more. There are other couples that are often referenced but are not often used today such as the Normal Hydrogen Electrode.

4.5.1. Three-Electrode Experiments:

In three electrode mode, the Reference lead is separated from the Counter and connected to a third electrode. This electrode is most often positioned so that it is measuring a point very close to the working electrode (which has both Working and Working Sense leads attached). A diagram of a 3-electrode cell setup can be seen in Figure 10.

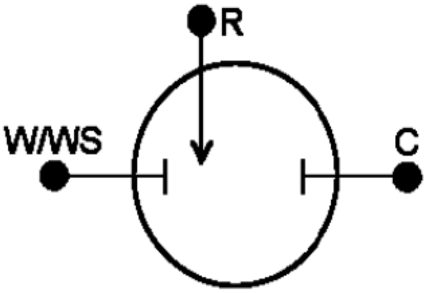


Fig. 10: 3-electrode cell setup

Three-electrode setups have a distinct experimental advantage over two electrode setups: they measure only one half of the cell. That is, the potential changes of the working electrode are measured independent of changes that may occur at the counter electrode. This isolation allows for a specific reaction to be studied with confidence and accuracy. Due to this reason, 3-electrode mode the most common setup used in electrochemical experimentation.

Chapter – 5

Major Components of the Project

5.1. Soil:

Soil is a thin layer of material on the Earth's surface in which plants have their roots. It is made up of many things, such as weathered rock and decayed plant and animal matter. Soil is formed over a long period of time.

Soil means different things to different people. Earth scientists see soil as mineral or organic material that is formed on Earth's surface by dynamic, complex processes. Engineers think of soil as material to build on and are concerned with moisture conditions and the ability of soil to become compacted and hold weight.

Soils are a mixture of different things; rocks, minerals, and dead, decaying plants and animals. Soil can be very different from one location to another, but generally consists of organic and inorganic materials, water and air. The inorganic materials are the rocks that have been broken down into smaller pieces. The size of the pieces varies. It may appear as pebbles, gravel, or as small as particles of sand or clay. The organic material is decaying living matter. This could be plants or animals that have died and decay until they become part of the soil. The amount of water in the soil is closely linked with the climate and other characteristics of the region. The amount of water in the soil is one thing that can affect the amount of air. Very wet soil like we would find in a wetland probably has very little air. The composition of the soil affects the plants and therefore the animals that can live there.

Depending upon the type of soil, there can be up to 5 different horizons. These are denoted by the letters O, A, B, C, and E. Not all soils will have these horizons, with some immature soils having none. Most soils have at least three of these (A, B, and C).

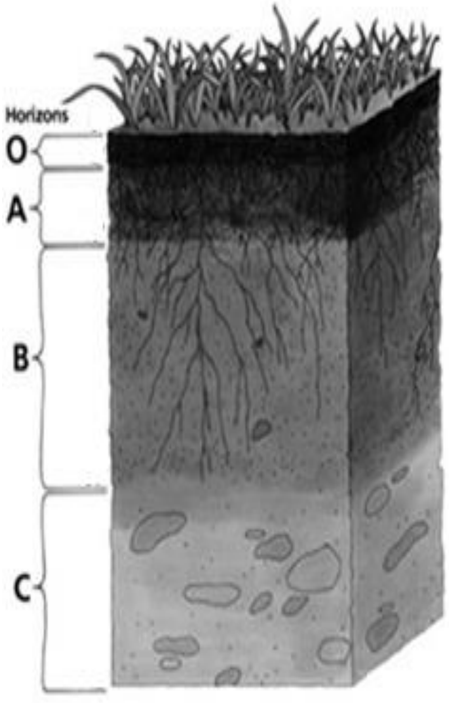


Fig.11: Layers of Soil

5.1.1. Soil Characteristics:

- **Color:**

The color of objects, including soils, can be determined by minor components. Generally, moist soils are darker than dry ones and the organic component also makes soils darker. Thus, surface soils tend to be darker than sub soil. Red, yellow and gray hues of sub soils reflect the oxidation and hydration states or iron oxides, which are reflective of predominant aeration and drainage characteristics in subsoil. Red and yellow hues are indicative of good drainage and aeration, critical for activity of aerobic organisms in soils. Mottled zones, splotches of one or more colors in a matrix of different color, often are indicative of a transition between well drained, aerated zones and poorly drained, poorly aerated ones. Gray hues indicate poor aeration. Soil color charts have been developed for the quantitative evaluation of colors.

- **Consistence:**

Consistence is a description of a soil's physical condition at various moisture contents as evidenced by the behavior of the soil to mechanical stress or manipulation. Descriptive adjectives such as hard, loose, friable, firm, plastic, and sticky are used for consistence. Soil consistence is of fundamental importance to the engineer who must move the material or compact it efficiently. The consistence of a soil is determined to a large extent by the texture of the soil, but is related also to other properties such as content of organic matter and type of clay minerals.

5.1.2. Physical Properties of Soil:

The physical properties of a soil are those characteristics which can be seen with the eye or felt between the thumb and fingers. They are the result of soil parent materials being acted upon by climatic factors (such as rainfall and temperature), and affected by topography (slope and direction, or aspect) and life forms (kind and amount, such as forest, grass, or soil animals) over a period of time. A change in any one of these influences usually results in a difference in the type of soil formed. Important physical properties of a soil are color, texture, structure, drainage, depth, and surface features (stoniness, slope, and erosion). Fertility is more easily changed than soil physical properties. These are:

- **Soil Texture:**

The size distribution of primary mineral particles, called *soil texture*, has a strong influence on the properties of a soil. Particles larger than 2 mm in diameter are considered inert. Particles which are smaller than 2 mm in diameter are divided into three broad categories based on size. Particles of 2 to 0.05 mm diameter are called sand; those of 0.05 to 0.002 mm diameter are silt; and the <0.002 mm particles are clay.

The texture of soils is usually expressed in terms of the percentages of sand, silt, and clay. To avoid quoting exact percentages, 12 textural classes have been defined. Each class, named to identify the size separate or separates having the dominant impact on properties, includes a range in size distribution that is consistent with a rather narrow range in soil behavior. The loam textural class contains soils whose properties are controlled equally by clay, silt and sand separates. Such soils tend to exhibit good balance between large and small pores; thus, movement of water, air and roots is easy and water retention is adequate. Soil texture, a stable and an easily determined soil characteristic, can be estimated by feeling and manipulating a moist sample, or it can be determined accurately by laboratory analysis. Soil horizons are sometimes separated on the basis of differences in texture.

Most surface soils fall into five general textural classes. Each class name indicates the size of the mineral particles that are dominant in the soil. Intermediate texture soils are called loams. Texture is determined in the field by rubbing moist-to-wet soil between the thumb and fingers. These observations can be checked in the laboratory by mechanical analysis which separates particles into clay, silt, and various-sized sand groups.

The sizes of these particles fall into the following sizes:

- sand is the coarsest (0.06 - 2 mm)
- silt is intermediate (0.002 - 0.06 mm)
- clay is the finest (<0.002 mm)

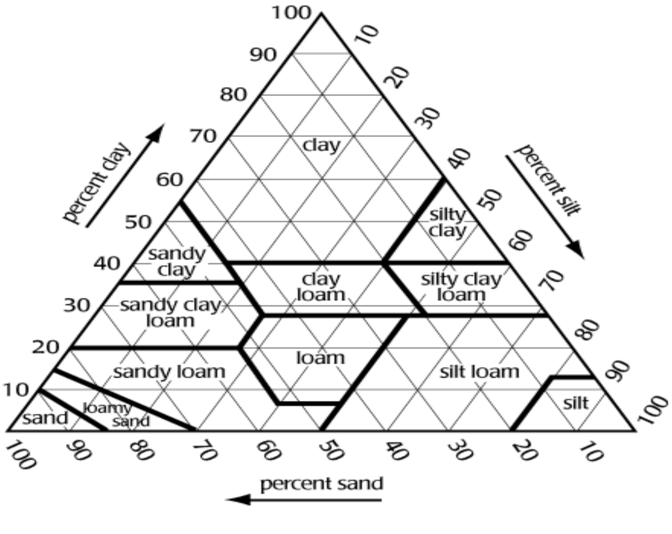


Fig. 12: Soil Texture Triangle

The sides of the soil texture triangle are scaled for the percentages of sand, silt, and clay. Clay percentages on the left side of the triangle are read from left to right across the triangle (dashed lines). Silt runs from the top to the bottom along the right side and is read from the upper right to lower left (light, dotted lines). The percentage of sand increases from right to left along the base of the triangle. Sand is read from the lower right towards the upper left portion of the triangle (bold, solid lines). The boundaries of the soil texture classes are highlighted in blue. The intersection of the three sizes on the triangle gives the texture class. For instance, if you have a soil with 20% clay, 60% silt, and 20% sand it falls in the "silt loam" class.

5.1.3. Chemical Properties of Soil:

- **Cation Exchange Capacity:**

Silicate clays and organic matter typically possess net negative charge because of cation substitutions in the crystalline structures of clay and the loss of hydrogen cations from functional groups of organic matter. Positively-charged cations are attracted to these negatively-charged particles, just as opposite poles of magnets attract one another. **Cation exchange** is the ability of soil clays and organic matter to adsorb and exchange cations with those in soil solution (water in soil pore space). A dynamic equilibrium exists between adsorbed cations and those in soil solution. Cation adsorption is reversible if other cations in soil solution are sufficiently concentrated to displace those attracted to the negative charge on clay and organic matter surfaces. The quantity of cation exchange is measured per unit of soil weight and is termed **cation exchange capacity**.

- **Soil pH:**

Soil pH is probably the most commonly measured soil chemical property and is also one of the more informative. Like the temperature of the human body, soil pH implies certain characteristics that might be associated with a soil. Since pH (the negative log of the hydrogen ion activity in solution) is an inverse, or negative, function, soil pH decreases as hydrogen ion, or acidity, increases in soil solution. Soil pH increases as acidity decreases.

A soil pH of 7 is considered neutral. Soil pH values greater than 7 signify alkaline conditions, whereas those with values less than 7 indicate acidic conditions. Soil pH typically ranges from

4 to 8.5, but can be as low as 2 in materials associated with pyrite oxidation and acid mine drainage. In comparison, the pH of a typical cola soft drink is about 3.

- **Soil Salinity:**

Soil salinity is the salt content of the soil. Salt affected soils are caused by excess accumulation of salts, typically most pronounced at the soil surface. Salts can be transported to the soil surface by capillary transport from a salt laden water table and then accumulate due to evaporation; they can also be concentrated in soils due to human activity. As soil salinity increases, salt effects can result in degradation of soils and vegetation. Salinization is a process that results from:

5.1.4. Macronutrients of Soil:

Nitrogen:

Much of the nitrogen reserve is stored in the soil as organic matter and most of this organic fraction is found in the upper soil horizons. At surface mines, the upper soil horizons are usually removed and stockpiled prior to disturbance. The storage of topsoil allows for relatively rapid conversion of organic nitrogen to soluble nitrate (NO3-) and is subject to leaching or conversion to nitrogen gas (denitrification) which volatilizes out of solution into the atmosphere. Thus, when stored topsoil is spread on a disturbed landscape, nitrogen reserves may be depleted or altered by several chemical and biological phenomena and the healthy cycling of nitrogen through the ecosystem inhibited or prevented. [42] Commonly, nitrogen fertilizer is land applied to reclaim mine sites where revegetation is desired. Tilling is generally not necessary to incorporate the nitrogen into the soil because of the leaching ability of nitrogen.

Nitrogen in soils can be in various different forms. Nitrogen is very dynamic and is constantly changing chemical species and concentrations. In most soils, nitrate is the common ionic form of plant-available nitrogen, but this element may also exist as ammonium (NH4+) or nitrite (NO2-) as well as other ions.

Phosphorus:

Phosphorous is usually plant-available in soil as inorganic phosphate ions (HPO42- and H2PO42-) and sometimes as soluble organic phosphorous. The HPO42- anion dominates in strongly acidic soils while the H2PO4- anion dominates in alkaline soils. Both anions are important in near-neutral soils. The major portion of the total soil phosphorous - 96% to 99% - is not plant-available. The bulk of the soil phosphorous exists in three general groups of compounds - namely, organic phosphorous, calcium-bound inorganic phosphorous, and iron- or aluminum-bound inorganic phosphorous. Most of these phosphorous groups have very low solubility and are not readily available for plant uptake. When soluble sources of phosphorous, such as fertilizers and manures, are added to soils, they are fixed and, in time, form highly insoluble compounds that are not plant available. Fixation reactions in soils may allow only small fractions (10% to 15%) of the phosphorous in fertilizers and manures to be taken up by plants in the year of application. Consequently, when budget allows, application of two to four times as much phosphorous than expected for plant uptake is common Unlike nitrogen, phosphorous has a low solubility and, therefore, it must be incorporated into the soil with a plow, disk, or chisel to ensure good soil-root-phosphorus contact. [43]

Potassium:

The original sources of potassium are the primary minerals, such as micas (biotite and muscovite) and potassium feldspar (orthoclase and microcline). As these minerals weather, the potassium becomes more available as readily exchangeable and soluble potassium which can be absorbed by plants roots. At any one time, most soil potassium is in primary minerals and non exchangeable forms. In relatively fertile soils, the release of potassium from these forms to the exchangeable and soil solution forms that plants can use directly may be sufficiently rapid to keep plants supplied with enough potassium for optimum growth. Conversely, in relatively non fertile soils, the levels of exchangeable and solution potassium may have to be supplemented by outside sources, such as chemical fertilizers, poultry manure, or wood ashes. Without these additions, the supply of available potassium will likely be depleted over a period of years and the productivity of the soil will likewise decline. [44]

5.2. Electrochemical Impedance Spectroscopy:

Electrical impedance spectroscopy (EIS) is a technology for obtaining bio-impedance measurements at different frequencies. For an electrochemical cell, such as those used for bioparticle detection, the impedance between the electrodes is determined by physical properties of the electrodes and the electrolyte, as well as chemical interactions between the metal and the ions in the electrolyte. Each of these current conducting processes can be represented by an electronic element. [45]

EIS involves measurements and analysis of materials in which ionic conduction strongly predominates. Examples of such materials are solid and liquid electrolytes, fused salts, ionically conducting glasses and polymers, and nonstoichiometric ionically bonded single crystals, where conduction can involve motion of ion vacancies and interstitials. EIS is also valuable in the study of fuel cells, rechargeable batteries, and corrosion. [46]

It is a powerful diagnostic tool that you can use to characterize limitations and improve the performance of fuel cells. There are three fundamental sources of voltage loss in fuel cells: charge transfer activation or “kinetic” losses, ion and electron transport or “ohmic” losses, and concentration or “mass transfer” losses. Among other factors, EIS is an experimental technique that can be used to separate and quantify these sources of polarization. By applying physically-sound equivalent circuit models wherein physiochemical processes occurring within the fuel cell are represented by a network of resistors, capacitors and inductors, you can extract meaningful qualitative and quantitative information regarding the sources of impedance within the fuel cell. EIS is useful for research and development of new materials and electrode structures, as well as for product verification and quality assurance in manufacturing operations.

During an impedance measurement, a frequency response analyzer (FRA) is used to impose a small amplitude AC signal to the fuel cell via the load. The AC voltage and current response of the fuel cell is analyzed by the FRA to determine the resistive, capacitive and inductive behavior the impedance of the cell at that particular frequency. Physicochemical processes occurring within the cell – electron & ion transport, gas & solid phase reactant transport, heterogeneous reactions, etc. have different characteristic time-constants and therefore are exhibited at different AC frequencies. When conducted over a broad range of frequencies, impedance spectroscopy can be used to identify and quantify the impedance associated with these various processes.

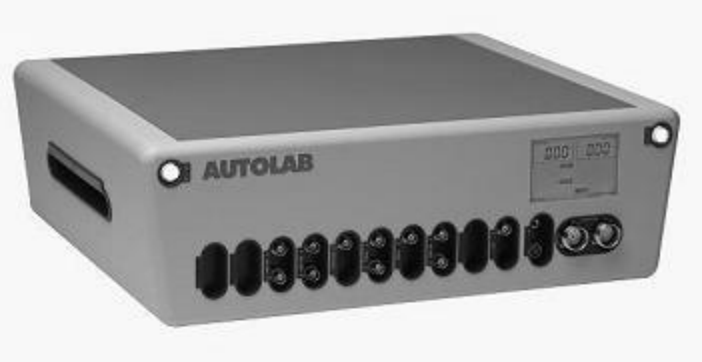


Fig. 13: Auto Lab Electrochemical Impedance Spectroscopy Setup

The impedance measurements are advantageous in comparison to potentiometry and especially voltammetry, owing to the potential experimental simplicity and the reduction of the response times. EIS has been applied for qualitative discrimination of mineral water, tea, coffee and red wine. [47]

5.3. Unscrambler:

The main purpose of The Unscrambler is to provide tools which can help one to analyze multivariate data. This involves finding variations, co-variations and other internal relationships in data matrices (tables). One can also use The Unscrambler to conduct a design of experiments (DOE) resulting in well planned experiments to achieve the maximum possible information.

The following are the basic types of problems that can be solved using The Unscrambler:

- Design experiments, analyze effects and find optima using the Design Experiment Wizard
- Reformat and preprocess data to enhance future analyses
- Find relevant variation in one data matrix (X)
- Find relationships between two data matrices (X and Y)
- Validate multivariate models with Uncertainty Testing
- Resolve unknown mixtures by finding the number of pure components and estimating their concentration profiles and spectra

- Predict the unknown values of a response variable
- Classify unknown samples into various possible categories

5.3.1. Make Well-Designed Experimental Plans:

Choosing samples carefully increases the chance of extracting useful information from data. Furthermore, being able to actively experiment with the variables also increases the chance of extracting relationships. The critical part is deciding which variables to change, which intervals to use for this variation, and the pattern of the experimental points.

The purpose of *experimental design* is to generate experimental data that enable one to determine which *design variables* (X) have an influence on the *response variables* (Y), in order to understand the interactions between the design variables and thus determine the optimum conditions. Of course, it is equally important to do this with a minimum number of experiments to reduce costs. An experimental design program should offer appropriate design methods and encourage good experimental practice, i.e. allow one to perform few but useful experiments which span the important variations.

- *Screening designs* (e.g. fractional, full factorial and Plackett-Burman) are used to find out which design variables have an effect on the responses and are suitable for collection of data spanning all important variations.
- *Optimization designs* (e.g. central composite, Box-Behnken) aim to find the optimum conditions for a process and generate nonlinear (quadratic) models. They generate data tables that describe relationships in more detail, and are usually used to refine a model, i.e. after the initial screening has been performed.

There are several methods for analysis of experimental designs. The Unscrambler uses Multiple Linear Regression (MLR) as its default methods for *orthogonal* designs. For non-orthogonal designs, or when the levels of a design cannot be reached, The Unscrambler allows the use other methods, such as PCR or PLS, for this purpose.

5.3.2. Reformat, Transform And Plot Data:

Raw data may have a distribution that is not optimal for analysis. Background effects, measurements in different units, different variances in variables etc. may make it difficult for the methods to extract meaningful information. Preprocessing or transformations help in reducing the “noise” introduced by such effects.

Before applying transforms, it is important to look at the data from a slightly different point of view. Sorting samples or variables and transposing the data table are examples of such reformatting operations.

Whether the data have been reformatted and transformed or not, a quick plot may reveal more about the data than is to be seen with the naked eye on a mere collection of numbers. Various types of plots are available in The Unscrambler. They facilitate visual checks of individual variable distributions, allow one to study the correlation among two variables or examine samples as for example 3-D swarm of points or a 3-D landscape.

5.3.3. Study Variations Among One Group Of Variables:

A common problem is to determine which variables actually contribute to the variation seen in a given data matrix; i.e. to find answers to questions such as:

- “Which variables are necessary to describe the samples adequately?”
- “Which samples are similar to each other?”
- “Are there groups of samples in a particular data set?”
- “What is the meaning of these sample patterns?”

The Unscrambler finds this information by decomposing the data matrix into a structured part and a noise part, using a technique called *Principal Component Analysis (PCA)*.

5.3.4. Other Methods to Describe One Group of Variables:

Classical *descriptive statistics* are also available in The Unscrambler. Mean, standard deviation, minimum, maximum, median and quartiles provide an overview of the univariate distributions of variables, allowing for their comparison. In addition, the *correlation* matrix provides a summary of the co variations among variables.

In the case of instrumental measurements performed on samples representing mixtures of a few pure components at varying concentrations or at different stages of a process (such as chromatography), The Unscrambler offers a method for recovering the unknown concentrations, called Multivariate Curve Resolution (MCR).

5.3.5. Study Relations between Two Groups Of Variables:

Another common problem is establishing a *regression model* between two data matrices. For example, one may have a set of many inexpensive measurements (X) of properties of a set of different solutions (for example), and want to relate these measurements to the concentration of a particular compound (Y) in the solution. The concentrations of the particular compound are usually found using a reliable reference method.

In order to do this, it is necessary to find the relationship between the two data matrices. This task varies somewhat depending on whether the data have been generated using statistical experimental design or have simply been collected, more or less at random, from a given population (i.e. non-designed data).

5.3.6. How To Analyze Designed Data Matrices:

The variables in designed data tables (excluding mixture or D-optimal designs) are orthogonal. Traditional statistical methods such as ANOVA and MLR are well suited to make a regression model from orthogonal data tables.

5.3.7. How to Analyze Non-Designed Data Matrices:

The variables in non-designed data matrices are seldom orthogonal, but rather more or less collinear with each other. MLR will most likely fail in such circumstances, so the use of *projection techniques* such as PCR or PLS is recommended.

5.3.8. Validate Multivariate Models With Uncertainty Testing:

Whatever the purpose in multivariate modeling – explore, describe precisely, build a predictive model – *validation* is an important issue. Only a proper validation can ensure that the model results are not too highly dependent on some extreme samples, and that the predictive power of the regression model meets the experimental objectives.

5.3.9. Estimate New, Unknown Response Values:

A regression model can be used to predict new, i.e. unknown, Y-values. Prediction is a useful technique as it can replace costly and time consuming measurements. A typical example is the prediction of concentrations from absorbance spectra instead of direct measurements of them by, for example titration.

5.3.10. Classify Unknown Samples:

Classification simply means to find out whether new samples are similar to classes of samples that have been used to make models in the past. If a new sample fits a particular model well, it is said to be a member of that class. Classification can be done using several different techniques including SIMCA, LDA, SVM classification and PLS-DA.

5.3.11. Reveal Groups Of Samples:

Clustering attempts to group samples into 'k' clusters based on specific distance measurements.

In The Unscrambler, clustering can be applied to a data set using the K-Means algorithm, as well as using hierarchical clustering (HCA). Seven different types of distance measurements are provided along with popular algorithms, including Ward's method.

Overall, The Unscrambler is a complete, All-In-One Multivariate Data Analysis and Design of Experiment package, which can be used to investigate simple, through to extremely large and complex data tables, for most applications. It provides the analytical tools most commonly used and requested by most data analysts. The plug in architecture allows for the inclusion new transforms and methods as they become available and software validation has been greatly simplified as a result of this. The Unscrambler meets the data security requirements for regulated industries.

Chapter – 6

Results Analysis

6.1. Project Aim:

The aim of the project titled *'Development of Soil Nutrient Prediction Model Using Pattern Recognition Techniques'* is to develop a model which will be able to correctly predict the concentration of the macronutrients (Nitrogen, Phosphorus & Potassium) of the unknown soil sample. The success of this project will bring a major change in the agricultural sector as accurate measurements of soil macronutrients are needed for efficient agricultural production, including site-specific crop management (SSCM), where fertilizer nutrient application rates are adjusted spatially based on local requirements.

The success of the project is dependent on how well the model is trained using the sample data and how well the model is able to classify & predict the unknown soil sample. For this, the model is trained using the multivariate data analysis software *'Unscrambler X 10.2'*.

6.2. Experimental Details:

6.2.1. Material Used:

Electrode electrochemical impedance spectroscopy is proposed for the discrimination of soil samples. Impedance response of gold working electrode in the frequency range of 1 kHz to 100 kHz has been measured. For this, 3 electrode setup is used for taking the impedance reading of the samples. The 3 electrodes used are as follows:

1. Working Electrode
2. Counter Electrode
3. Reference Electrode

Working electrode is the designation for the electrode being studied. This is most often an inert material which will pass current to other species without being affected by that current. In our case, this electrode is Au (Gold).

The Counter electrode is the electrode in the cell that completes the current path. All electrochemistry experiments (with non-zero current) must have a working – counter pair. In most experiments the Counter is simply the current source/sink and so relatively inert materials like graphite or platinum are ideal, though not necessary. We are using separate Pt wire electrode as the counter electrode.

Reference electrodes are, as their name suggests, electrodes that serve as experimental reference points. Specifically they are a reference for the potential (sense) measurements. Reference electrodes should, therefore, hold a constant potential during testing, ideally one which is known on an absolute scale. We are using Ag/AgCl Wire as the reference electrode.

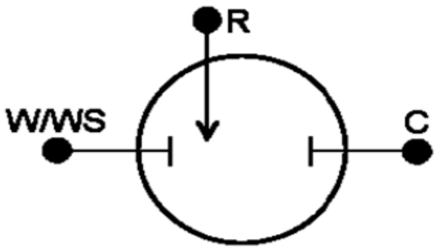


Fig.14: 3 Electrode Setup

6.2.2. Equipments Used:

CH Instruments, Electrochemical Workstation 660C was used to study the impedance behavior of the samples. Electrochemical impedance spectroscopy measurements are performed with the Auto lab instruments in combination with the FRA2 module, or FRA module in short.

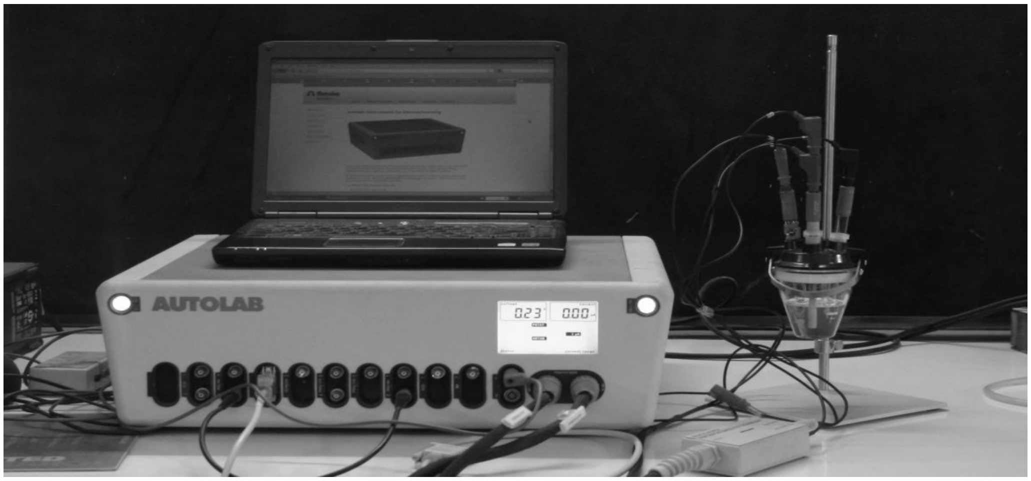


Fig. 15: Electrochemical Workstation 660C

Statistical Analysis software The Unscrambler (CAMO, Norway) was used for performing data analysis and identifying the patterns between the samples.

Project name: New Projec
Login: guest

Version: 10.2.0.80
UserID: guest@7FBAF56
Path Saved:

Fig. 16: Unscrambler 10.2X Opening Screen

6.2.3. Sample preparation:

The samples are prepared for three concentrations of Nitrogen, Potassium & Phosphate thus making a total of 27 (3x3x3) samples. The samples are made in de – ionized water. Nitrate used is- Magnesium nitrate, Phosphate is- Ammonium dihydrogen phosphate, Potassium is- Potassium chloride.

Table 1: Different Concentrations of Ions

IONS	CONCENTRATION		
Nitrate	N1: 600 ppm	N2: 2000 ppm	N3: 4000 ppm
Phosphate	P1: 800 ppm	P2: 2500 ppm	P3: 4500 ppm
Potassium	K1: 700 ppm	K2: 2500 ppm	K3: 4500 ppm

Depending upon this, there are following 27 samples:

Table 2: 27 Samples & their concentrations

N1_P1_K1	N1_P1_K2	N1_P1_K3
N1_P2_K1	N1_P2_K2	N1_P2_K3
N1_P3_K1	N1_P3_K2	N1_P3_K3
N2_P1_K1	N2_P1_K2	N2_P1_K3
N2_P2_K1	N2_P2_K2	N2_P2_K3
N2_P3_K1	N2_P3_K2	N2_P3_K3
N3_P1_K1	N3_P1_K2	N3_P1_K3
N3_P2_K1	N3_P2_K2	N3_P2_K3
N3_P3_K1	N3_P3_K2	N3_P3_K3

6.3. Experimental Setup:

The technique that is illustrated in the figure below is electrochemical impedance spectroscopy which is proposed for the discrimination of soil samples. Impedance response of gold working electrode in the frequency range of 1 kHz to 100 kHz has been measured.

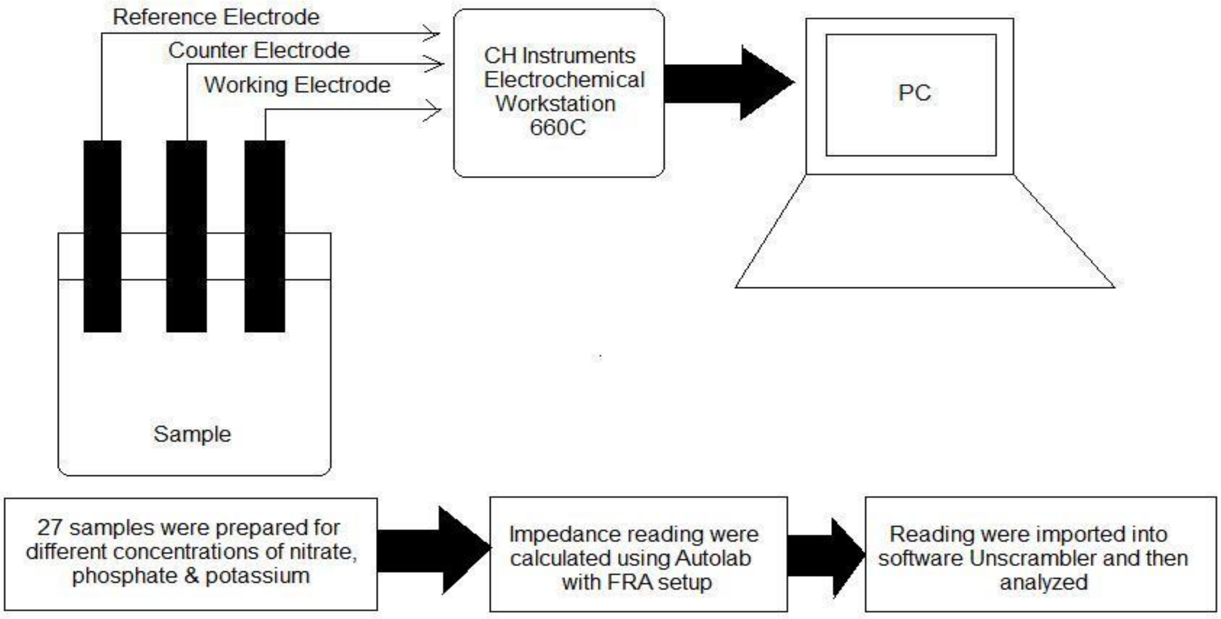

Reference Electrode
Counter Electrode
Working Electrode
CH Instruments Electrochemical Workstation 660C
PC
Sample
27 samples were prepared for different concentrations of nitrate, phosphate & potassium
Impedance reading were calculated using Autolab with FRA setup
Reading were imported into software Unscrambler and then analyzed

Fig. 17: Experimental Set Up

6.4. Results & Analysis:

6.4.1. Objective of the analysis:

The main objective of performing data analysis using the multivariate data analysis software *'Unscrambler 10.2X'* is to deduce following conclusions:

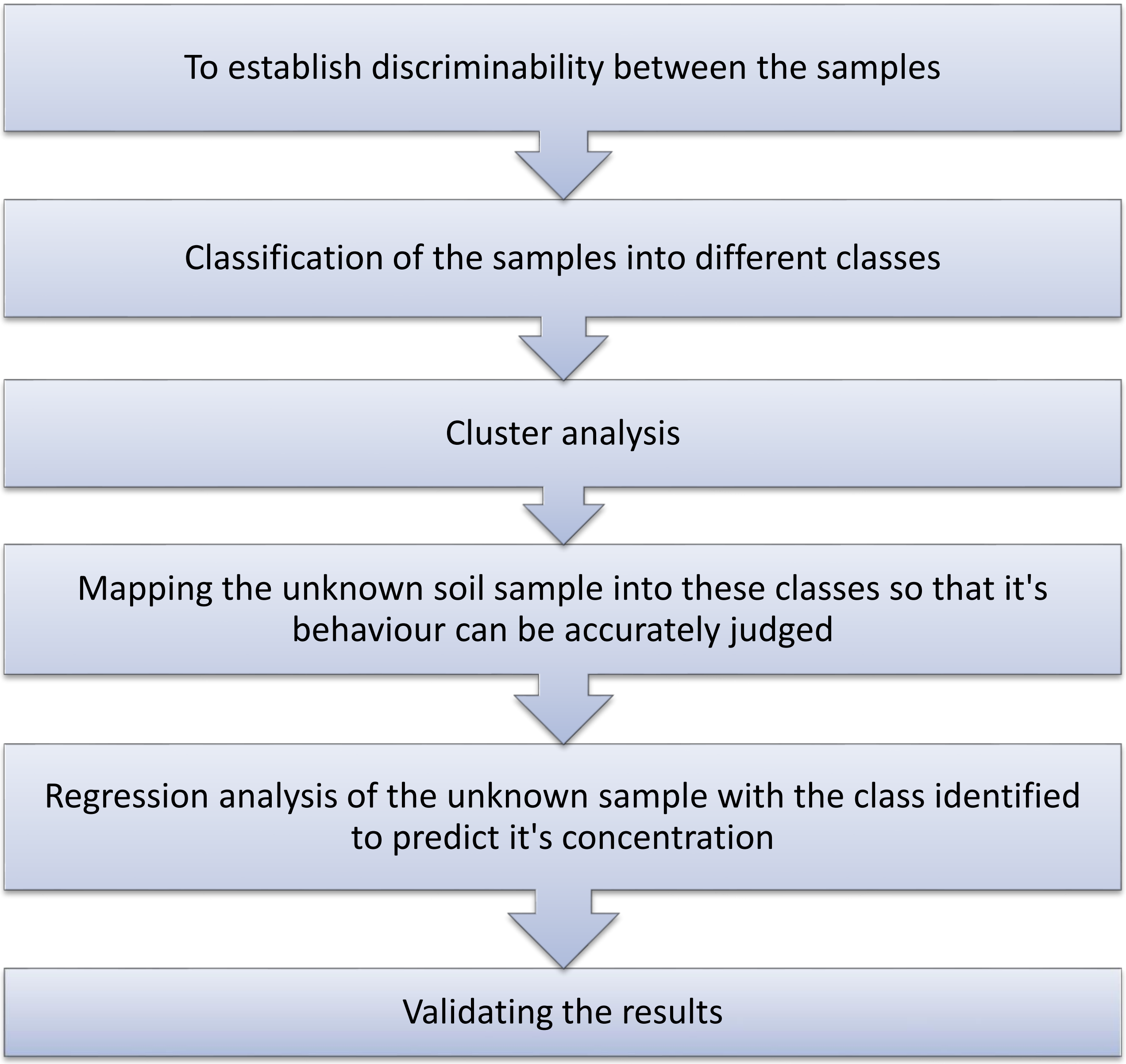

To establish discriminability between the samples
Classification of the samples into different classes
Cluster analysis
Mapping the unknown soil sample into these classes so that it's behaviour can be accurately judged
Regression analysis of the unknown sample with the class identified to predict it's concentration
Validating the results

Fig.18: Flow chart of the analysis performed

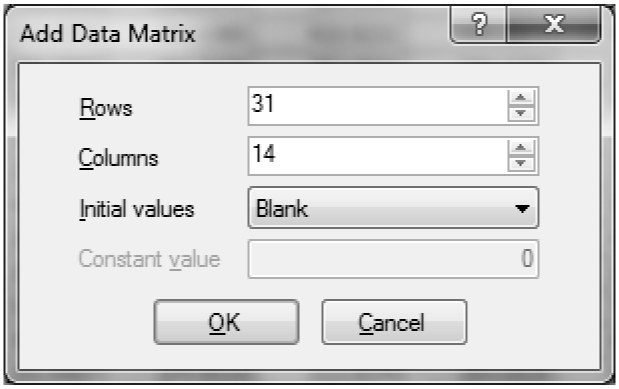
6.4.2. Data Matrix Design:

The data is taken for three combinations of nitrate, phosphate & potassium ions making it a total of 27 (3x3x3) samples. The readings are taken at a range of frequency 1 kHz to 100 kHz. 10 frequencies have been taken that cover the entire range. Hence, trained data set comprises of 270 readings.

The data matrix is defined as 31x14 2D matrix where the rows represent the 27 samples plus 4 unknown samples & columns represents the frequency range & the concentrations of the ions in that sample. A11 in the matrix represents the impedance reading (|z| = square root of squares of real impedance part & imaginary impedance part).

Rows are referred to as ‘samples’ & columns are referred as ‘variables’ that affect the behavior of samples.

The following dialog box appears to create the data matrix:



Add Data Matrix

Rows 31

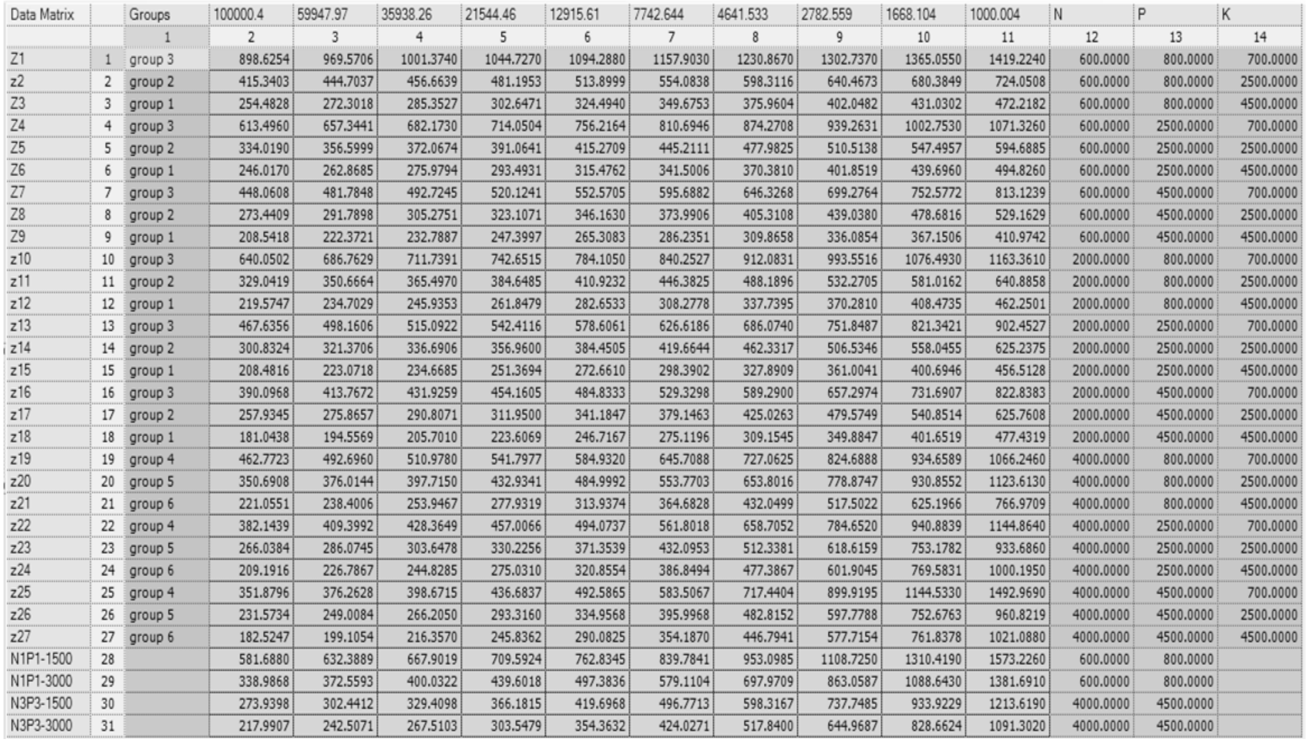
Columns 14

Initial values Blank

Constant value 0

Fig. 19: Dialog Box for creating Matrix

We are using following data matrix to perform the analysis:



Data Matrix		Groups	100000.4	59947.97	35938.26	21544.46	12915.61	7742.644	4641.533	2782.559	1668.104	1000.004	N	P	K
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Z1	1	group 3	898.6254	969.5706	1001.3740	1044.7270	1094.2880	1157.9030	1230.8670	1302.7370	1365.0550	1419.2240	600.0000	800.0000	700.0000
z2	2	group 2	415.3403	444.7037	456.6639	481.1953	513.8999	554.0838	598.3116	640.4673	680.3849	724.0508	600.0000	800.0000	2500.0000
Z3	3	group 1	254.4828	272.3018	285.3527	302.6471	324.4940	349.6753	375.9604	402.0482	431.0302	472.2182	600.0000	800.0000	4500.0000
Z4	4	group 3	613.4960	657.3441	682.1730	714.0504	756.2164	810.6946	874.2708	939.2631	1002.7530	1071.3260	600.0000	2500.0000	700.0000
Z5	5	group 2	334.0190	356.5999	372.0674	391.0641	415.2709	445.2111	477.9825	510.5138	547.4957	594.6885	600.0000	2500.0000	2500.0000
Z6	6	group 1	246.0170	262.8685	275.9794	293.4931	315.4762	341.5006	370.3810	401.8519	439.6960	494.8260	600.0000	2500.0000	4500.0000
Z7	7	group 3	448.0608	481.7848	492.7245	520.1241	552.5705	595.6882	646.3268	699.2764	752.5772	813.1239	600.0000	4500.0000	700.0000
Z8	8	group 2	273.4409	291.7898	305.2751	323.1071	346.1630	373.9906	405.3108	439.0380	478.6816	529.1629	600.0000	4500.0000	2500.0000
Z9	9	group 1	208.5418	222.3721	232.7887	247.3997	265.3083	286.2351	309.8658	336.0854	367.1506	410.9742	600.0000	4500.0000	4500.0000
z10	10	group 3	640.0502	686.7629	711.7391	742.6515	784.1050	840.2527	912.0831	993.5516	1076.4930	1163.3610	2000.0000	800.0000	700.0000
z11	11	group 2	329.0419	350.6664	365.4970	384.6485	410.9232	446.3825	488.1896	532.2705	581.0162	640.8858	2000.0000	800.0000	2500.0000
z12	12	group 1	219.5747	234.7029	245.9353	261.8479	282.6533	308.2778	337.7395	370.2810	408.4735	462.2501	2000.0000	800.0000	4500.0000
z13	13	group 3	467.6356	498.1606	515.0922	542.4116	578.6061	626.6186	686.0740	751.8487	821.3421	902.4527	2000.0000	2500.0000	700.0000
z14	14	group 2	300.8324	321.3706	336.6906	356.9600	384.4505	419.6644	462.3317	506.5346	558.0455	625.2375	2000.0000	2500.0000	2500.0000
z15	15	group 1	208.4816	223.0718	234.6685	251.3694	272.6610	298.3902	327.8909	361.0041	400.6946	456.5128	2000.0000	2500.0000	4500.0000
z16	16	group 3	390.0968	413.7672	431.9259	454.1605	484.8333	529.3298	589.2900	657.2974	731.6907	822.8383	2000.0000	4500.0000	700.0000
z17	17	group 2	257.9345	275.8657	290.8071	311.9500	341.1847	379.1463	425.0263	479.5749	540.8514	625.7608	2000.0000	4500.0000	2500.0000
z18	18	group 1	181.0438	194.5569	205.7010	223.6069	246.7167	275.1196	309.1545	349.8847	401.6519	477.4319	2000.0000	4500.0000	4500.0000
z19	19	group 4	462.7723	492.6960	510.9780	541.7977	584.9320	645.7088	727.0625	824.6888	934.6589	1066.2460	4000.0000	800.0000	700.0000
z20	20	group 5	350.6908	376.0144	397.7150	432.9341	484.9992	553.7703	653.8016	778.8747	930.8552	1123.6130	4000.0000	800.0000	2500.0000
z21	21	group 6	221.0551	238.4006	253.9467	277.9319	313.9374	364.6828	432.0499	517.5022	625.1966	766.9709	4000.0000	800.0000	4500.0000
z22	22	group 4	382.1439	409.3992	428.3649	457.0066	494.0737	561.8018	658.7052	784.6520	940.8839	1144.8640	4000.0000	2500.0000	700.0000
z23	23	group 5	266.0384	286.0745	303.6478	330.2256	371.3539	432.0953	512.3381	618.6159	753.1782	933.6860	4000.0000	2500.0000	2500.0000
z24	24	group 6	209.1916	226.7867	244.8285	275.0310	320.8554	386.8494	477.3867	601.9045	769.5831	1000.1950	4000.0000	2500.0000	4500.0000
z25	25	group 4	351.8796	376.2628	398.6715	436.6837	492.5865	583.5067	717.4404	899.9195	1144.5330	1492.9690	4000.0000	4500.0000	700.0000
z26	26	group 5	231.5734	249.0084	266.2050	293.3160	334.9568	395.9968	482.8152	597.7788	752.6763	960.8219	4000.0000	4500.0000	2500.0000
z27	27	group 6	182.5247	199.1054	216.3570	245.8362	290.0825	354.1870	446.7941	577.7154	761.8378	1021.0880	4000.0000	4500.0000	4500.0000
N1P1-1500	28		581.6880	632.3889	667.9019	709.5924	762.8345	839.7841	953.0985	1108.7250	1310.4190	1573.2260	600.0000	800.0000	
N1P1-3000	29		338.9868	372.5593	400.0322	439.6018	497.3836	579.1104	697.9709	863.0587	1088.6430	1381.6910	600.0000	800.0000	
N3P3-1500	30		273.9398	302.4412	329.4098	366.1815	419.6968	496.7713	598.3167	737.7485	933.9229	1213.6190	4000.0000	4500.0000	
N3P3-3000	31		217.9907	242.5071	267.5103	303.5479	354.3632	424.0271	517.8400	644.9687	828.6624	1091.3020	4000.0000	4500.0000	

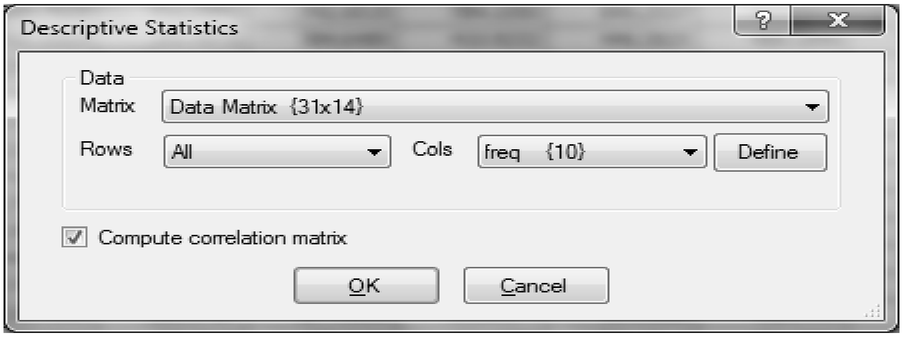
Fig: Data Matrix

6.4.3. Descriptive Statistics:

The Descriptive Statistics option provides some simple and effective plotting tools for gaining an overview of small to medium sized data sets.

Step 1: Tasks → Analyze → Descriptive Statistics

Step 2: Following dialog box appears: Rows depict the samples & the columns will be the 10 frequencies which affect the impedance readings:



Matrix: Data Matrix {31x14}

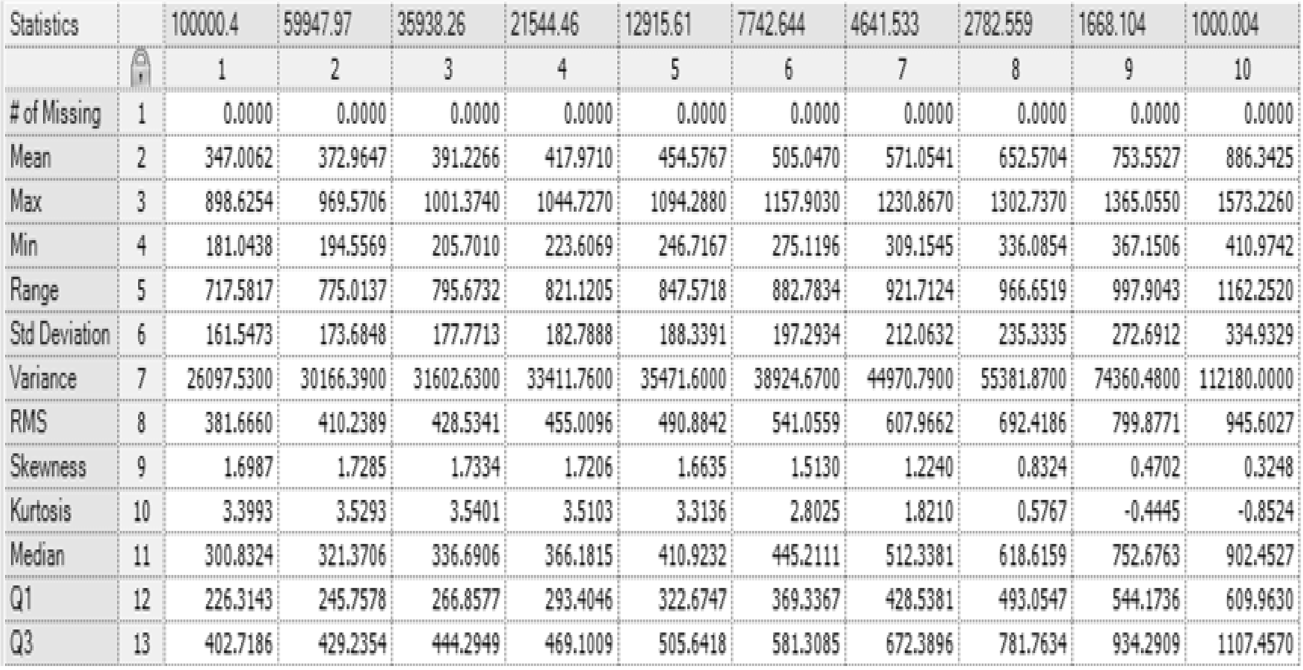
Rows: All

Cols: freq {10}

☑ Compute correlation matrix

Fig. : Dialog box to perform descriptive statistics

Step 3: Following results are obtained:



Statistics		100000.4	59947.97	35938.26	21544.46	12915.61	7742.644	4641.533	2782.559	1668.104	1000.004
	🔒	1	2	3	4	5	6	7	8	9	10
# of Missing	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Mean	2	347.0062	372.9647	391.2266	417.9710	454.5767	505.0470	571.0541	652.5704	753.5527	886.3425
Max	3	898.6254	969.5706	1001.3740	1044.7270	1094.2880	1157.9030	1230.8670	1302.7370	1365.0550	1573.2260
Min	4	181.0438	194.5569	205.7010	223.6069	246.7167	275.1196	309.1545	336.0854	367.1506	410.9742
Range	5	717.5817	775.0137	795.6732	821.1205	847.5718	882.7834	921.7124	966.6519	997.9043	1162.2520
Std Deviation	6	161.5473	173.6848	177.7713	182.7888	188.3391	197.2934	212.0632	235.3335	272.6912	334.9329
Variance	7	26097.5300	30166.3900	31602.6300	33411.7600	35471.6000	38924.6700	44970.7900	55381.8700	74360.4800	112180.0000
RMS	8	381.6660	410.2389	428.5341	455.0096	490.8842	541.0559	607.9662	692.4186	799.8771	945.6027
Skewness	9	1.6987	1.7285	1.7334	1.7206	1.6635	1.5130	1.2240	0.8324	0.4702	0.3248
Kurtosis	10	3.3993	3.5293	3.5401	3.5103	3.3136	2.8025	1.8210	0.5767	-0.4445	-0.8524
Median	11	300.8324	321.3706	336.6906	366.1815	410.9232	445.2111	512.3381	618.6159	752.6763	902.4527
Q1	12	226.3143	245.7578	266.8577	293.4046	322.6747	369.3367	428.5381	493.0547	544.1736	609.9630
Q3	13	402.7186	429.2354	444.2949	469.1009	505.6418	581.3085	672.3896	781.7634	934.2909	1107.4570

Fig. : Descriptive statistics of the data

Descriptive statistics help us to perform parametric statistics of the data set. By parametric statistics, it is inferred that the samples under investigation come from a population with a known underlying distribution, typically a *normal distribution*. A *Normal Distribution* is one where the population (or sample) values are symmetrically distributed around this mean value and the variance describes the width of the distribution. The normal distribution is therefore fully characterized by the two parameters, the mean and a measure of spread known as the *Standard Deviation*. Parametric statistics are sensitive to the underlying parameters, which in the case of a normal distribution are:

- The *Mean*, or the central tendency of the samples: this is the value where the distribution of sample values tends to congregate around a central value,
- The *Variance* or the spread of the samples around the mean.
- The *Standard Deviation* is a measure of spread, given in the same units as the original observations.
- *Max*: this tells us the maximum value under that particular variable.

- *Min*: this tells us the minimum value under that particular variable.
- *Range*: The Range of a data set is defined as the highest observed value minus the lowest observed value in a data set. It is a non-parametric method of describing dispersion and should be used instead of the standard deviation when the number of observations is less than 5.
- The *Skewness* of a distribution is a measure of its asymmetry and is referred to as the third central moment of the distribution. The degree of this asymmetry is determined by the coefficient of skewness.

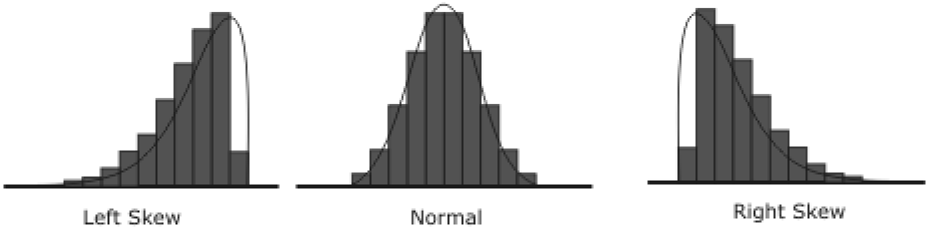


Fig. : Skewness

- The *Kurtosis* of a distribution is a different type of departure from normality compared to skewness. It describes the extent of the degree of flatness (or alternatively, the peakedness) of the center of a distribution. A value of the coefficient of kurtosis of around 0 indicates that the distribution is normal. When it is greater than 0, it indicates that there are more observations around the mean, i.e. the distribution is peaked. If the coefficient is less than 0, this indicates that the curve is flatter than normal.
- *Quartiles:* The Median represents the point in a data set that splits it into two equal parts. Quartiles take this idea further by splitting the data into four equal parts. These parts are labeled Q1, Q2 and Q3 respectively and Q2 represents the median. Another important measure in statistics is the *Interquartile Range* (IQR). The IQR is defined by the relationship

6.5. To Establish Discriminability Between The Samples:

The discriminability of the sample needs to be explained in terms of 3 factors i.e. ions, concentration & frequencies. These factors are being varied in different samples; hence they will provide us with a firm basis to discriminate the samples correctly.

- IONS: We are taking samples that have 3 ions in it i.e. the nitrate ion, the phosphate ion & the potassium ion. Hence, it is important to understand the effect of each ion on the sample correctly.
- CONCENTRATION: The samples are prepared for three different concentrations of each ion as explained in table 1. So effects of low, medium & high concentration needs to be identified correctly.
- FREQUENCY: The frequencies have a major effect on the impedance values of each sample. Hence, it should also be studied carefully.

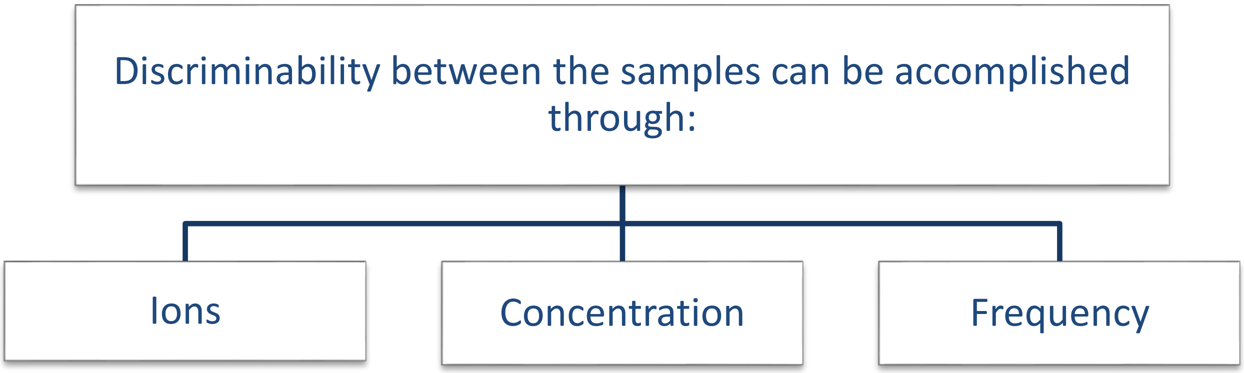

Discriminability between the samples can be accomplished through:
Ions
Concentration
Frequency

Fig: Discriminability between the samples

The discriminability will be demonstrated using a tool known as 'Principal Component analysis' or PCA.

6.5.1. PCA Plots:

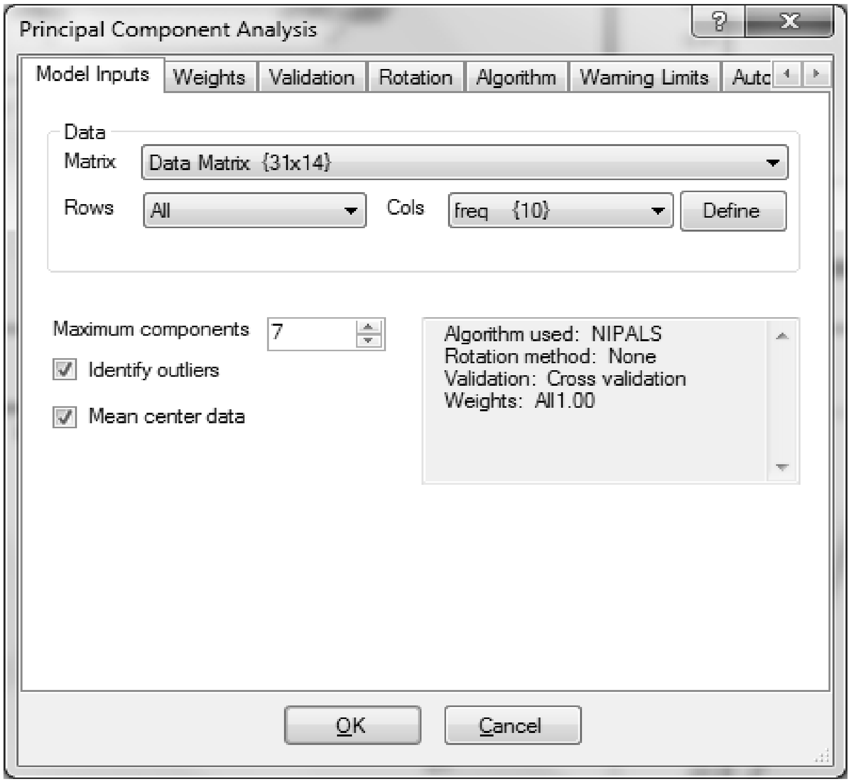
PCA is a statistical technique for the reduction of input data dimension and is largely used for feature extraction. It captures the relevant information in a set of input data providing a lower dimension, but informative representation of the original data. It sequentially creates a set of principal components from the original data. The first principal component (PC1) maps the maximum variance and information of the input data followed by the other principal components (PC2, PC3 and so on) in descending order of the variance. Generally a good discrimination is mapped by the first two principal components, i.e. PC1 and PC2. Statistical software The Unscrambler 10.2 has been used for the PCA.

6.5.2. Steps to do Principal Component Analysis:

Step 1: Tasks → Analyze → Principal Component Analysis

Step 2: In the *Model Inputs* tab, select a matrix to be analyzed from the Matrix drop-down list. The matrix comprises of all the samples & the columns represent the 10 frequencies.

The **Mean Center** check box allows a user to subtract the column means from every variable before analysis. The **Identify Outliers** check box allows a user to identify potential outliers based on parameters set up in the **Warning Limits** tab.



Matrix: Data Matrix {31x14}

Rows: All

Cols: freq {10}

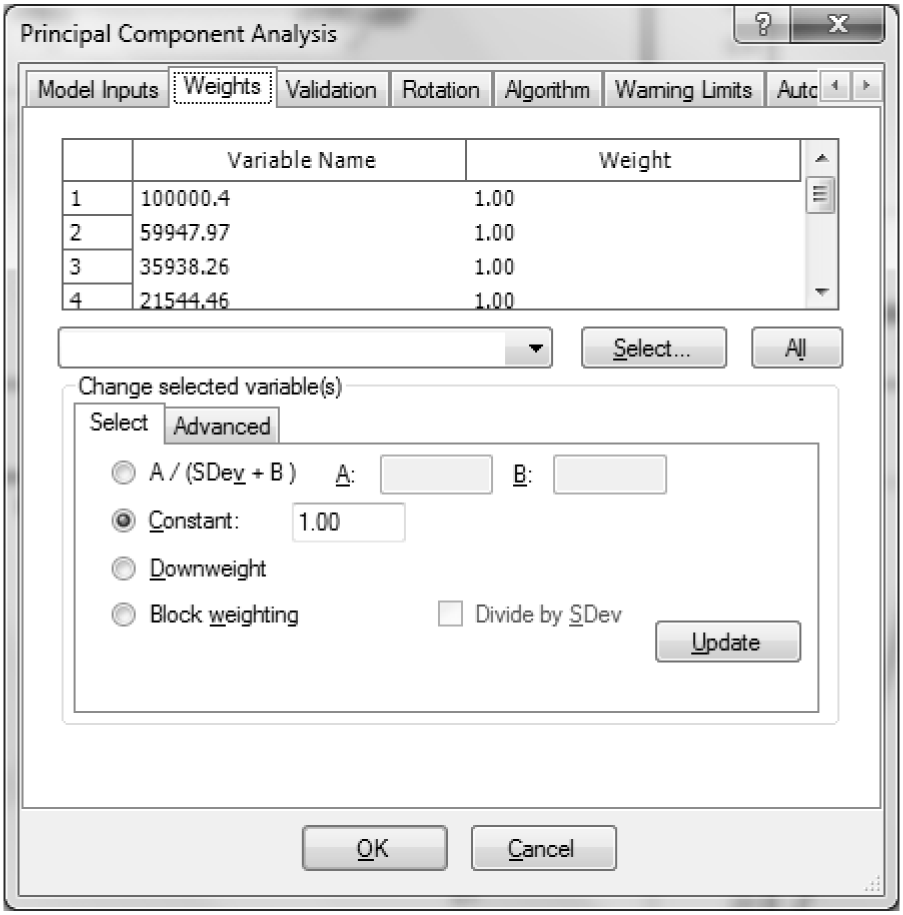
Maximum components: 7

☑ Identify outliers

☑ Mean center data

Algorithm used: NIPALS
Rotation method: None
Validation: Cross validation
Weights: All1.00

Step 3: If the analysis calls for variables to be weighted for making realistic comparisons to each other (particularly useful for process and sensory data), click on the **Weights** tab and the following dialog box will appear. **Constant** allows the weighting of selected variables by predefined constant values.



	Variable Name	Weight
1	100000.4	1.00
2	59947.97	1.00
3	35938.26	1.00
4	21544.46	1.00

○ A / (SDev + B) A: B:

● Constant: 1.00

○ Downweight

○ Block weighting □ Divide by SDev

Step 4: The next step in the PCA modeling process is to choose a suitable validation method from the **Validation** tab. The following dialog box will appear.

Cross Validation:

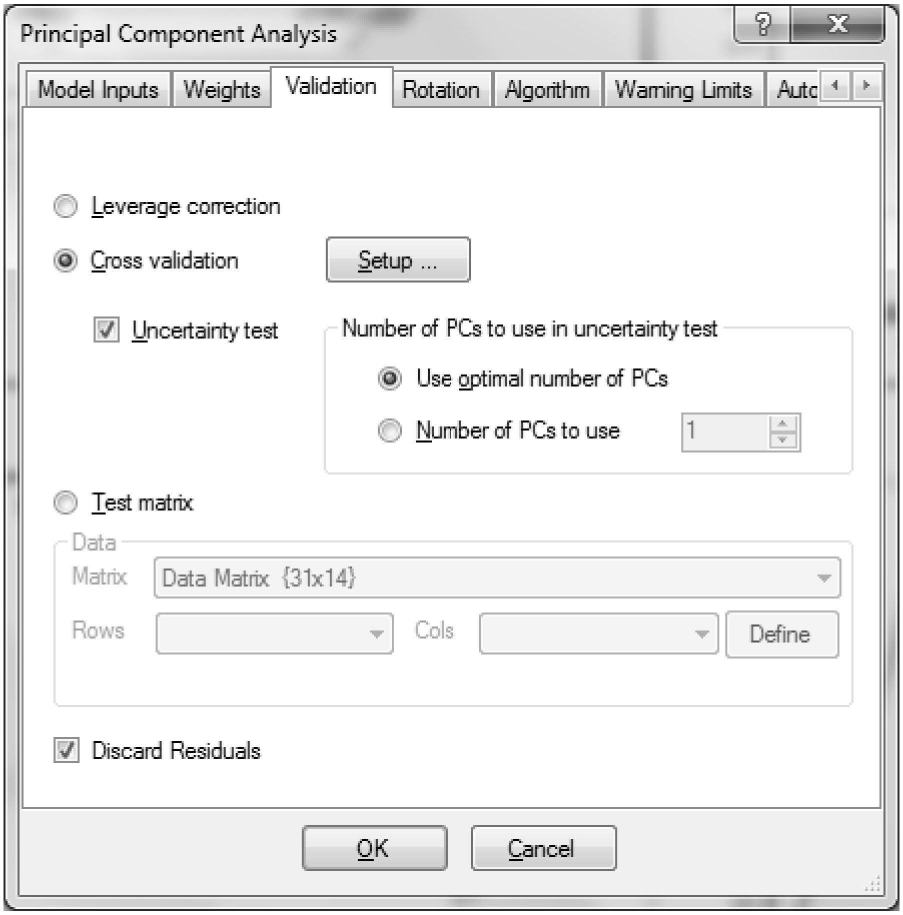
This method is used when either there are not enough samples available to make a separate test set, or for simulating the effects of different validation test cases, e.g. systematically leaving samples out vs. randomly leaving samples out, etc.

Uncertainty Test:

This method can be used to determine the significance of variables, when using cross validation, by applying the Marten’s Uncertainty Test. Check the **Uncertainty Test** box and the options available are to use the optimal number of PCs found in a model, or define the number of PCs to use for the test.

Discard Residuals option:

In The Unscrambler X all results from the modeling are stored to have the maximum flexibility in plotting any result matrix in any way to make the right decision regarding outliers, interpretation of the model etc. However, as the size of data matrices become large, the residual matrices use a lot of available memory and disk space, resulting in the size of the Unscrambler project becoming large and sometime unmanageable.



○ Leverage correction

● Cross validation

☑ Uncertainty test

Number of PCs to use in uncertainty test

● Use optimal number of PCs

○ Number of PCs to use 1

○ Test matrix

Data

Matrix Data Matrix {31x14}

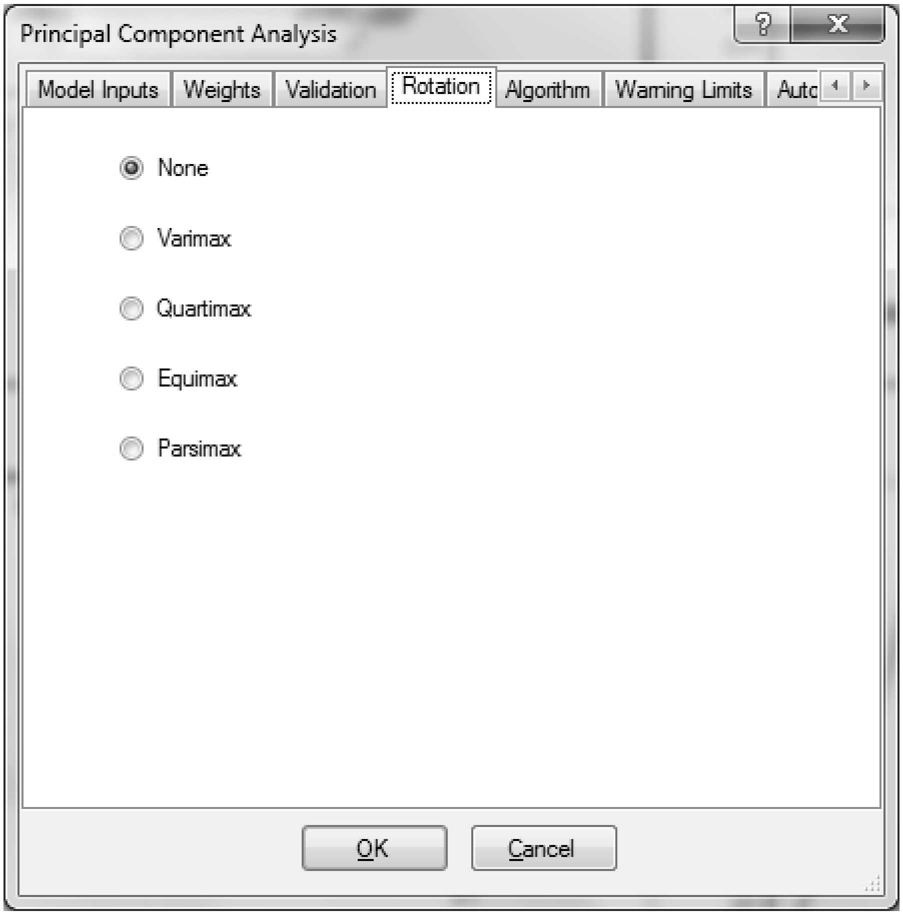
Rows

Cols

☑ Discard Residuals

Step 5: **PCA rotation**

The **Rotation** tab allows a user to apply rotation methods such as Varimax to a PCA model. But we are not applying any rotation in this.



● None

○ Varimax

○ Quartimax

○ Equimax

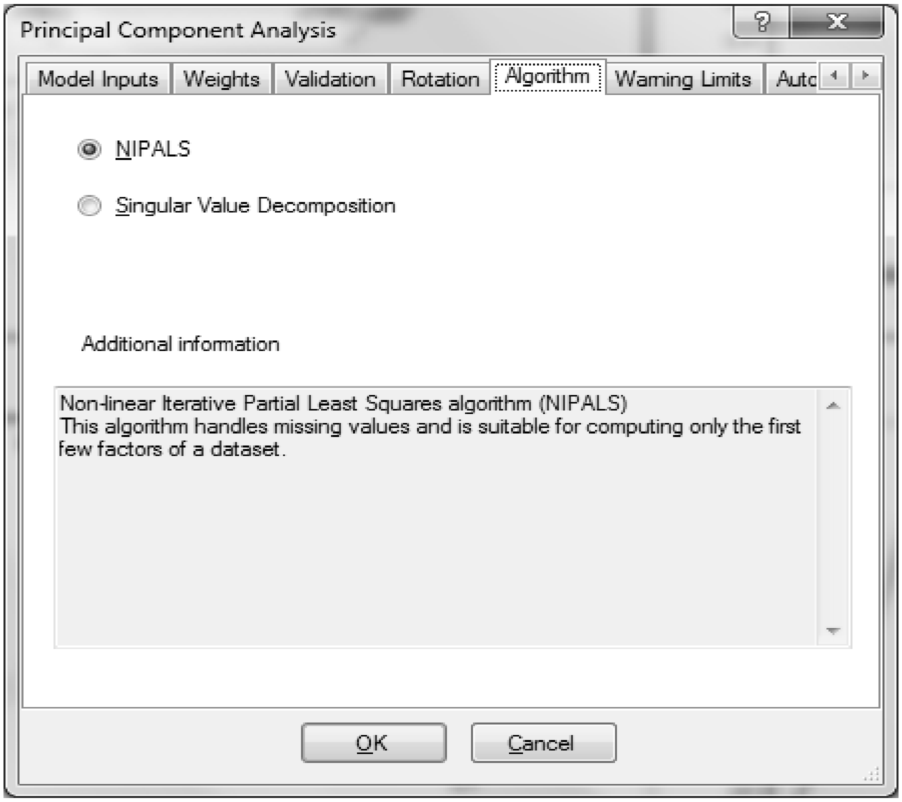
○ Parsimax

Step 6: The **Algorithms** tab gives a choice between the Non-linear Iterative Partial Least Squares (NIPALS) algorithm, used when missing values are present in the data set and the Singular Value Decomposition (SVD) algorithm, used for smaller data sets with no missing values present.

The algorithm we are using is:

NIPALS:

NIPALS stands for Non-linear Iterative Partial Least Squares. This algorithm handles missing values and is suitable for computing only the first few components of a large data set. This method however accumulates errors that can become large in higher principal components. The NIPALS algorithm calculates one principal component at a time and it handles missing values well, whereas the SVD algorithm calculates all of the principal components in one calculation, but does not handle missing values.



Principal Component Analysis

● NIPALS

○ Singular Value Decomposition

Additional information

Non-linear Iterative Partial Least Squares algorithm (NIPALS)
This algorithm handles missing values and is suitable for computing only the first few factors of a dataset.

6.5.3. PCA RESULTS:

Plot 1:

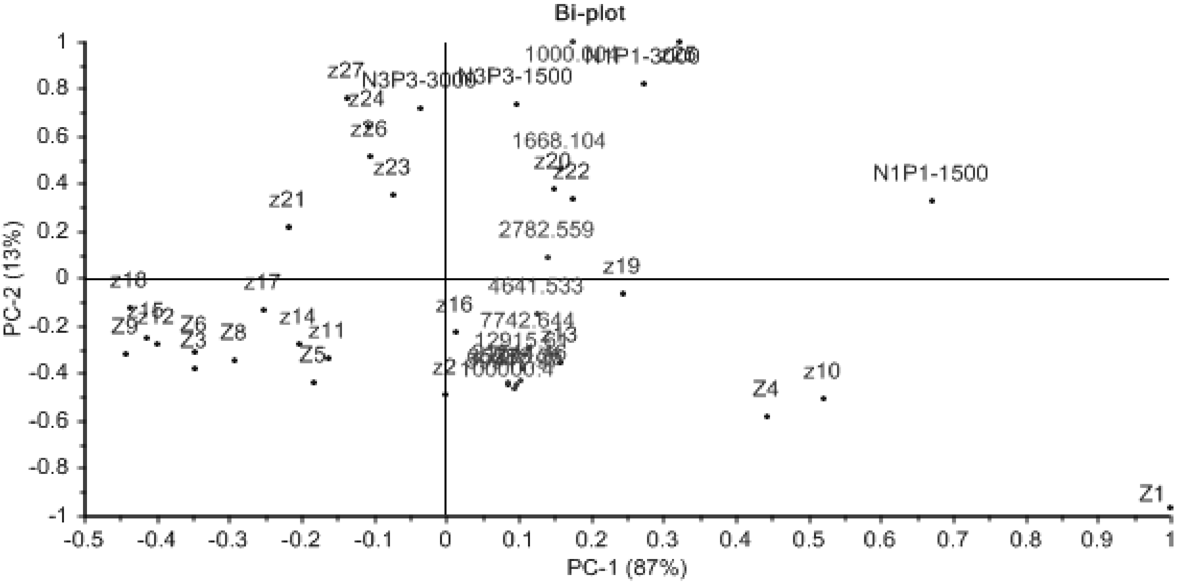


Fig: PCA biplot of samples with frequencies

Plot 2:

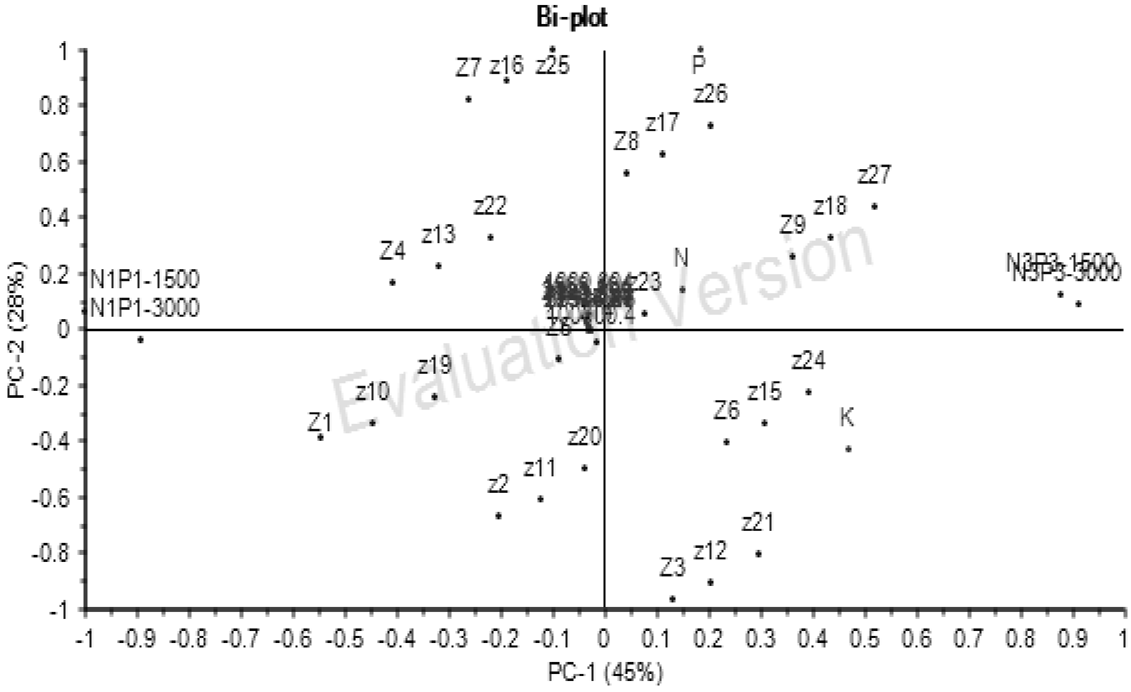


Fig. : PCA plot of samples with frequencies & concentrations

6.5.4. Interpreting PCA Plots:

The above two plots explain the discriminability between the samples in terms of ions, concentration & frequency.

Plot 1:

In plot 1, PC-1 accounts for 87% of the total explained variance & PC-2 accounts for 13% of the total explained variance. Thus combined, they explain 100% of the total variance. Hence, we need to consider only 2 PC components here. Moreover, we can identify six recognisable groups which are as follows:

Group 1: z3, z6, z9, z12, z15, z18

Group 2: z2, z5, z8, z11, z14, z17

Group 3: z1, z4, z7, z10, z13, z16

Group 4: z19, z22, z25

Group 5: z20, z23, z26

Group 6: z21, z24, z27

These groups form the basis of our classes for the classification. The groups are classified on the basis of concentration & ions. The group members have certain similarities. The noticeable thing about group 1 is that they have high concentration of potassium ions i.e. 4500 ppm and in group 2, the concentration of potassium ion is 2500 ppm & 700 ppm in group 1. So, we can say that variation of the concentration of the potassium ion can be seen along the PC-1 component.

Rest of the groups are distinguished on the basis of nitrate as well as potassium ions. Group 4 to 6 has high concentration of nitrate ions and the variation in nitrate ion concentration is established through the PC-2 component. These groups also have the same pattern for the potassium ion.

Lower n middle nitrate ion concentrations are accurately predicted by PC1 (which defines 87% of the total variance) & the high nitrate ion concentration samples are more correctly predicted by PC2 (which explains 13 % of the total explained variance). High concentration of nitrate ions is negatively correlated with low n middle concentration of nitrate ions since they lie in opposite quadrants.

Plot 2:

This is a bi plot made for all the samples with the frequencies & the concentration of the samples. In this case, PC-1 accounts for 45% of the total explained variance & PC-2 explains 28% of the total explained variance, hence combined the two PC components explain 73% of the total explained variance.

Since we are taking both the factors here, concentration & frequency, it shows that concentration & the ions play a much important role than the frequencies. It’s because of the reason that frequencies lie in the origin of both PC-1 & PC-2 & hence are not explained well by any of the two principal components. So, we can say that when both the ions & frequencies are concerned, the ions dominate the frequencies to discriminate the samples.

Here, also we find three groups n these groups are discriminated by phosphate ions.

Group 1: z1, z2, z3, z10, z11, z12, z19, z20, z21

Group 2: z4, z5, z6, z13, z14, z15, z22, z23, z24

Group 3: z7, z8, z9, z16, z17, z18, z25, z26, z27

Each group member has certain similarities and in this PCA plot discrimination is achieved by the phosphate ion. In group 1, the concentration of phosphate ion is low i.e. 800 ppm. For group 2, the concentration of phosphate is 2500 ppm & that for group 3; its value is 4500 ppm.

These groups can be further subdivided into groups which are discriminated by the nitrate levels.

Group 1 has further 3 subgroups:

G11: z1, z2, z3	conc. N1= 600 ppm n P1= 800 ppm
G12: z10, z11, z12	conc. N2= 2000 ppm n P1=800 ppm
G13: z19, z20, z21	conc. N3= 4000 ppm n P1= 800 ppm

Similarly For group 2:

G21: z4, z5, z6	conc. N1= 600 ppm n P2= 2500 ppm
G22: z13, z14, z15	conc. N2= 2000 ppm n P2=2500 ppm
G23: z22, z23, z24	conc. N3= 4000 ppm n P2= 2500 ppm

Similarly for group 3:

G31: z7, z8, z9	conc. N1= 600 ppm n P3= 4500 ppm
G32: z16, z17, z18	conc. N2= 2000 ppm n P3=4500 ppm
G33: z25, z26, z27	conc. N3= 4000 ppm n P3= 4500 ppm

Hence here the discriminability is achieved through the three ions i.e. phosphate ions, potassium ions & the nitrate ions.

Effect of Frequency:

The 10 frequencies show a parabolic path. The minima of the parabolic path occur at 50 kHz. Apart from the frequency at 100 kHz, the frequencies follows a pattern with the frequency with the highest value is at the bottom and the frequency with the least value is at the top. The three frequencies (1000 Hz, 1668 Hz, and 2782 Hz) are positively correlated with the samples that have high concentration of nitrate. We can imply that the low frequencies explain sample that have high concentration of nitrate ions i.e. conc. of nitrate = 4000 ppm.

Similarly the frequencies that has higher values (100 kHz, 5o kHz, 35 kHz, 21 kHz, 12 kHz, 8 kHz, 5 kHz) explains the samples that has low concentration of nitrate ions i.e. conc. of nitrate in these samples are 600 ppm and 2000 ppm.

So two set of frequencies can be obtained:

S1: 1000 Hz, 1668 Hz, and 2782 Hz

S2: 100 kHz, 5o kHz, 35 kHz, 21 kHz, 12 kHz, 8 kHz, 5 kHz

We can see that group 1 and group 2 mentioned above are negatively correlated to the s1 frequencies and groups 5 & 6 are negatively correlated to the high frequency band i.e. s2. Group 2 is defined well by the frequency band s2 as it is positively correlated to each other. Similarly group 4 is defined well by the set of frequencies s1 as both lie in the same quadrant.

6.6. Classification of Samples into These Classes:

Based on the analysis conducted in the previous section with the PCA plots, we have identified six groups which are as follows:

Class 1: S3, S6, S9, S12, S15, S18

Class 2: S5, S8, S11, S14, S17, S20

Class 3: S1, S4, S7, S10, S13, S16

Class 4: S19, S22, S25

Class 5: S20, S23, S26

Class 6: S21, S24, S27

Now the unknown soil samples will be mapped to these classes.

6.7. Mapping the Unknown into These Classes:

6.7.1. Support vector classification:

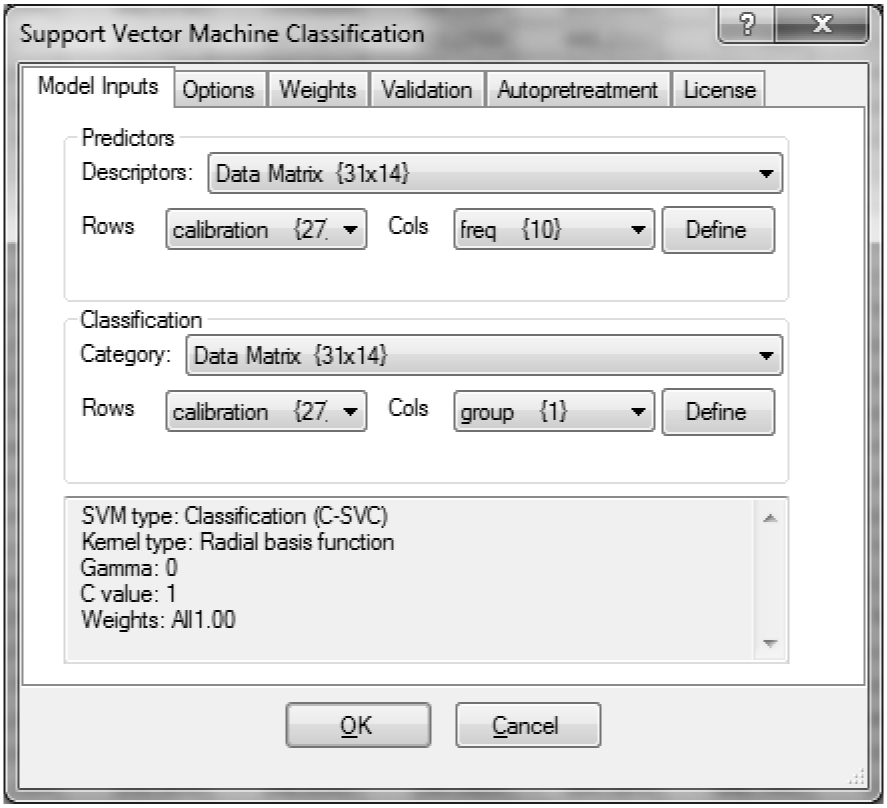
SVM is a classification method based on statistical learning wherein a function that describes a hyper plane for optimal separation of classes is determined. As the linear function is not always able to model such a separation, data are mapped into a new feature space and a dual representation is used with the data objects represented by their dot product. A kernel function is used to map from the original space to the feature space, and can be of many forms, thus providing the ability to handle nonlinear classification cases. The kernels can be viewed as a mapping of nonlinear data to a higher dimensional feature space, while providing a computation shortcut by allowing linear algorithms to work with higher dimensional feature space. The support vector is defined as the reduced training data from the kernel.

SVM classification is a supervised method of classification. The data used for SVM must have a data matrix which includes a single category variable defining which classes are to be discriminated by the model. The X and Y matrices must have the same number of rows (samples) for SVM classification, and not have any missing data. The Y matrix must contain a single column of category variables. The X data must be numerical, and not contain any missing data.

Steps to perform Support Vector Classification:

Step 1: Define the data matrix where the predictors defines the matrix or samples (training set) that are to classified and classification defines the group or classes into which the calibrated samples needs to be classified.

In our case, the predictors are the 27 samples that needs to be classified & the classification are the six classes into which the samples needs to be classified.

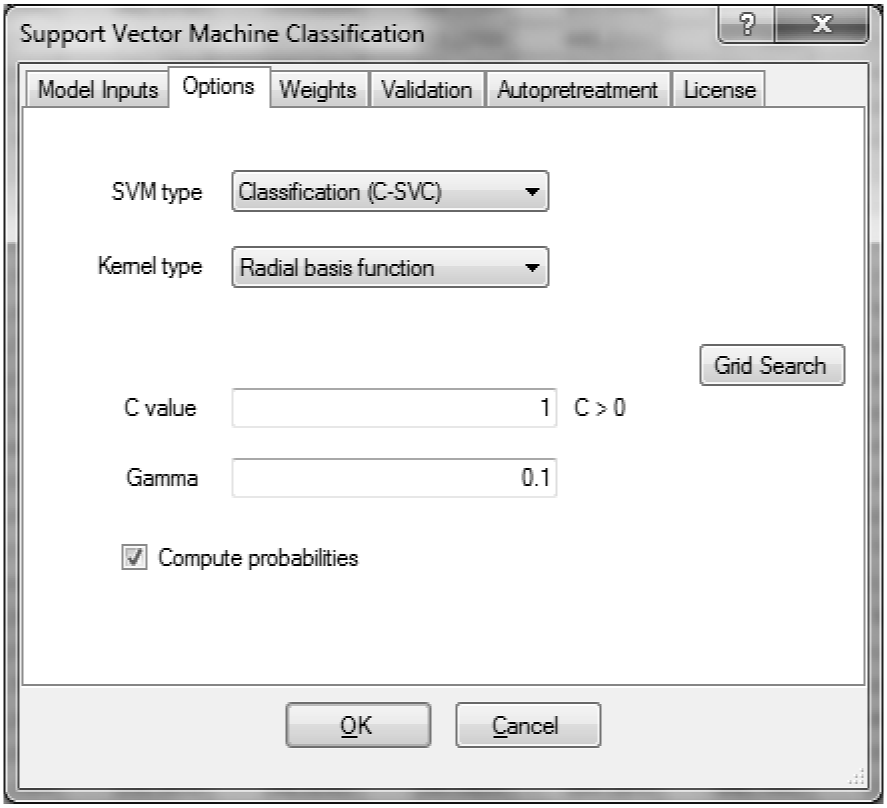


Predictors
Descriptors: Data Matrix {31x14}
Rows calibration {27 Cols freq {10}

Classification
Category: Data Matrix {31x14}
Rows calibration {27 Cols group {1}

SVM type: Classification (C-SVC)
Kernel type: Radial basis function
Gamma: 0
C value: 1
Weights: All1.00

Step 2: Here one can choose the SVM type of classification to use, either C-SVC or nu-SVM, from the drop-down list next to SVM type. The kernel type to be used to determine the hyper plane that best separates the classes can be selected from the following types from the drop-down list. The default setting of Radial basis function is the simplest, and can model complex data. The C-SVM has an input parameter named C, which is a capacity factor (also called penalty factor), a measure of the robustness of the model. C must be greater than 0.



SVM type: Classification (C-SVC)

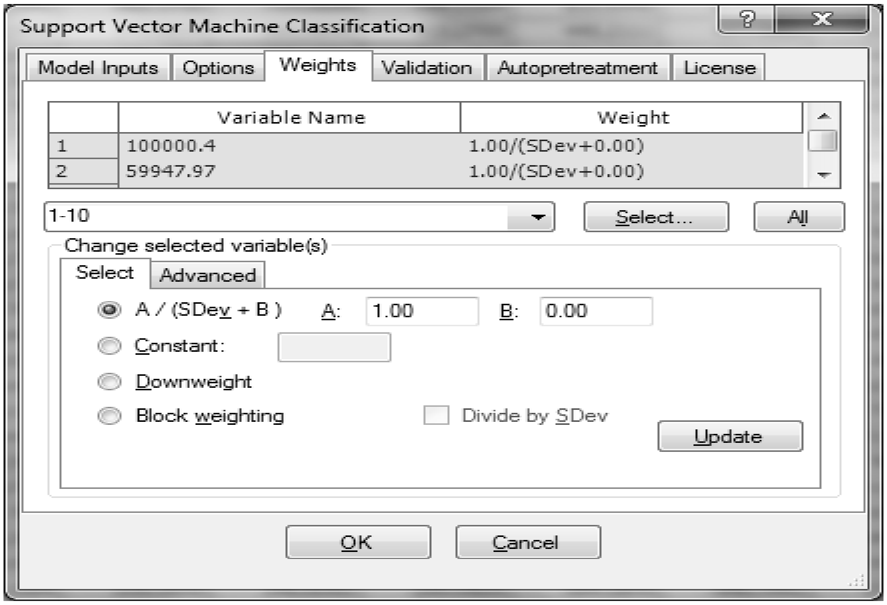
Kernel type: Radial basis function

C value: 1 C > 0

Gamma: 0.1

☑ Compute probabilities

Step 3: If the analysis calls for variables to be weighted for making realistic comparisons to each other (particularly useful for process and sensory data), click on the **Weights** tab and the following dialog box will appear. **Constant** allows the weighting of selected variables by predefined constant values.



	Variable Name	Weight
1	100000.4	1.00/(SDev+0.00)
2	59947.97	1.00/(SDev+0.00)

1-10

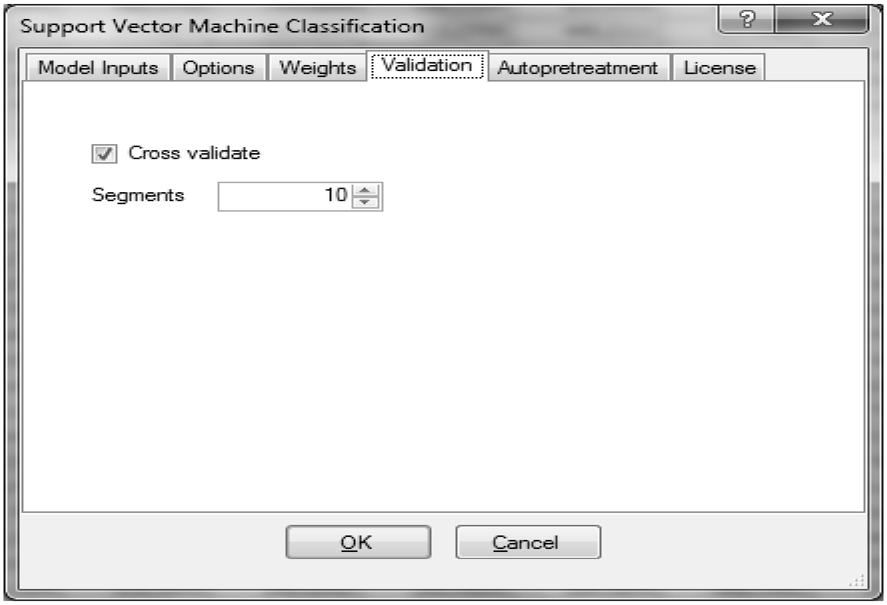
● A / (SDev + B) A: 1.00 B: 0.00

○ Constant:

○ Downweight

○ Block weighting □ Divide by SDev

Step 4: Validation is an important part of any method applied in modeling data. Settings for the Validation of the SVM are set under the Validation tab as shown below. First select to cross validate the model by checking the check box. The number of segments to use can be chosen in the **segments** entry. Cross validation is helpful in model development but should not be a replacement for full model validation using a test set.



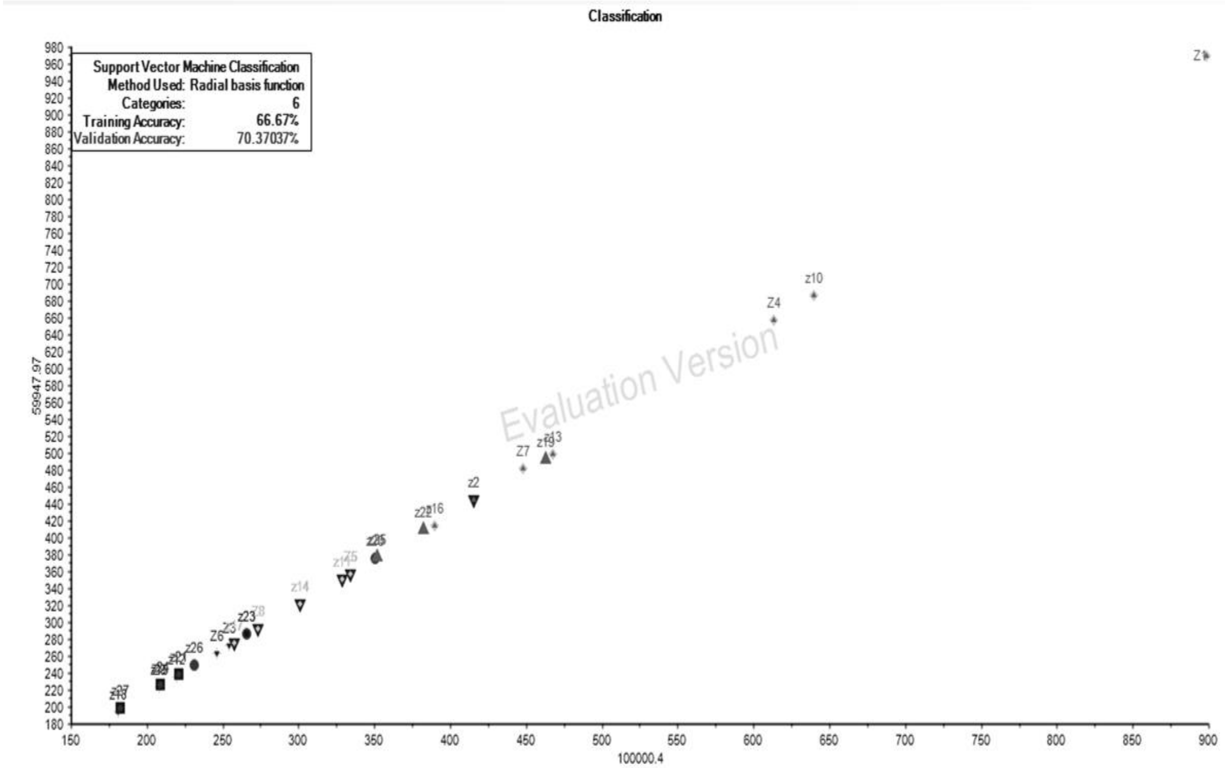
☑ Cross validate

Segments 10

6.7.2. Interpreting the Results of SVM:

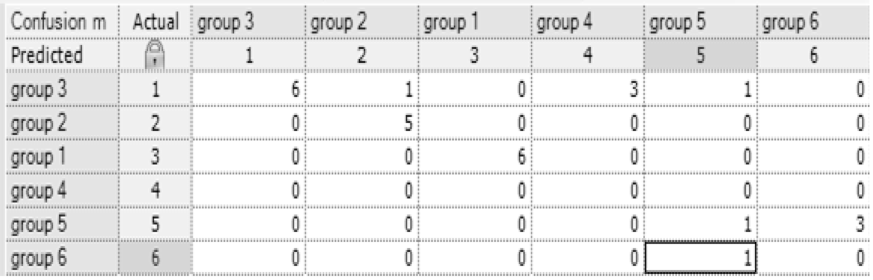
The samples are classified using SVM and the training accuracy is 67%. It means that the samples are classified up to 67% accurately. Validation percentage tells us how well the samples are validated or predicted accurately. The percentage for training accuracy can be improved by applying various other pretreatments like normalization techniques.

The graph given below shows the plot of the samples as per their impedance values at different frequencies. The different color coding tells us the group various samples belong to.



6.7.3. Confusion Matrix:

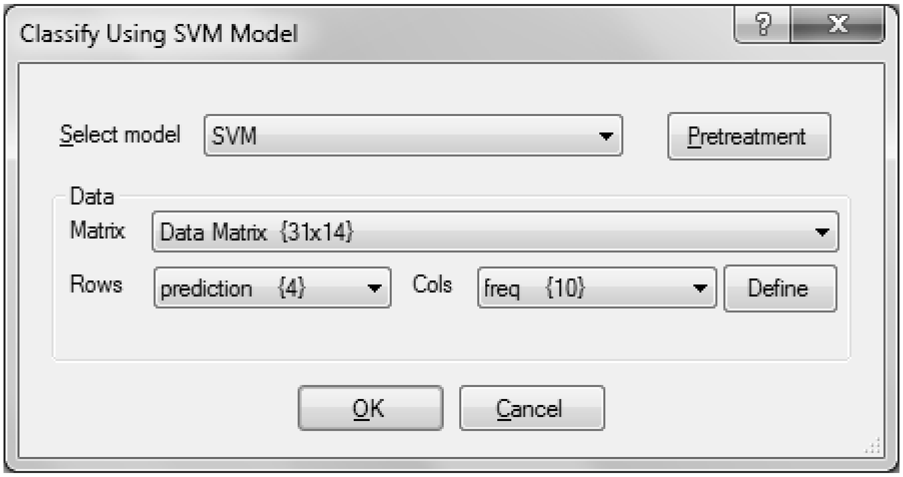
The confusion matrix is a matrix used for visualization for classification results from supervised methods such as support vector machine classification or linear discriminant analysis classification. It carries information about the predicted and actual classifications of samples, with each row showing the instances in a predicted class, and each column representing the instances in an actual class.



Confusion m	Actual	group 3	group 2	group 1	group 4	group 5	group 6
Predicted		1	2	3	4	5	6
group 3	1	6	1	0	3	1	0
group 2	2	0	5	0	0	0	0
group 1	3	0	0	6	0	0	0
group 4	4	0	0	0	0	0	0
group 5	5	0	0	0	0	1	3
group 6	6	0	0	0	0	1	0

6.7.4. Classifying the Unknown test samples:

We need to select the model that is created in the previous section i.e. SVM. The rows should be the rows or samples that need to be classified and the columns are the frequencies at which the impedance readings are taken.



Classify Using SVM Model

Select model: SVM

Data

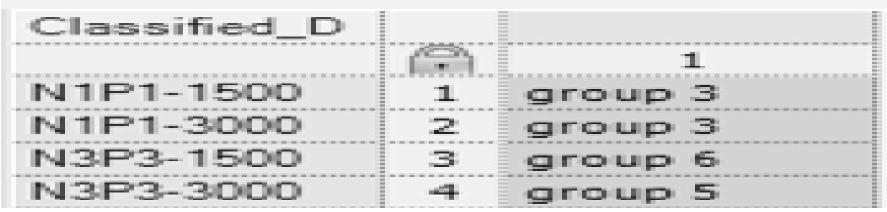
Matrix: Data Matrix {31x14}

Rows: prediction {4}

Cols: freq {10}

Classified Results:

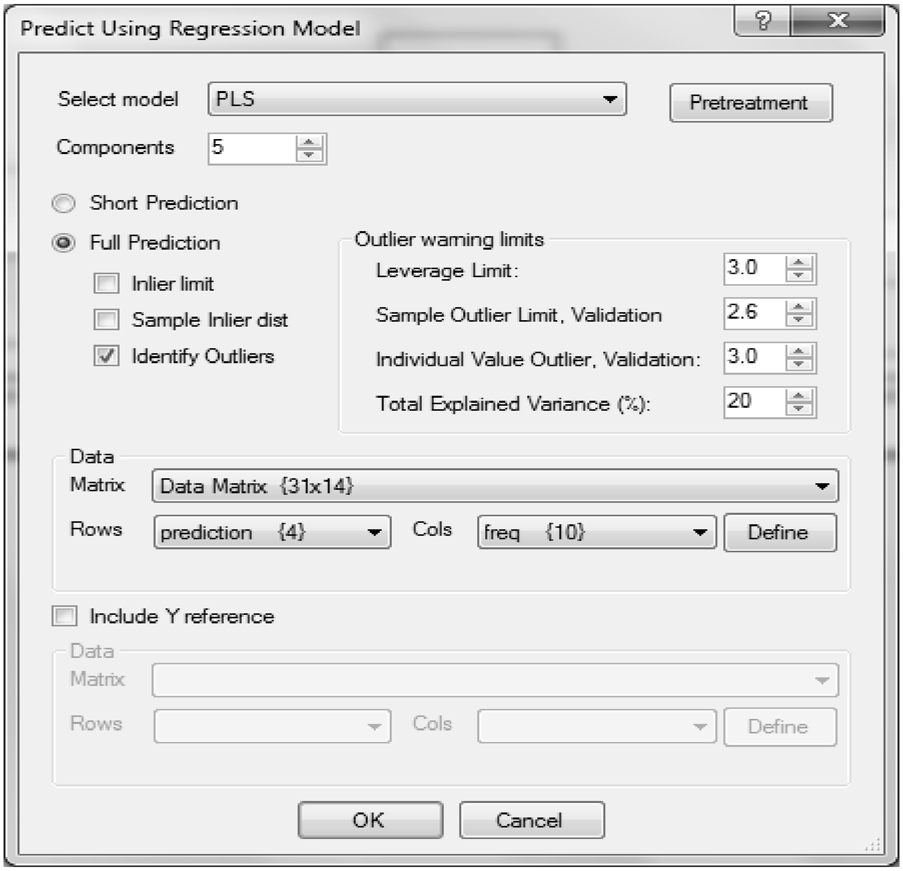
The results of the classified matrix are true for the three unknown samples i.e. N1P1-1500, N3P3-1500 & N3P3-3000. Since these are the laboratory test samples, their concentrations were known and the system is able to validate 75% of the results. The result is incorrect for the N1P1-3000 which may be due to the manual errors that might have happened during the experimentation times.



Classified_D		
		1
N1P1-1500	1	group 3
N1P1-3000	2	group 3
N3P3-1500	3	group 6
N3P3-3000	4	group 5

6.8. Regression analysis of the unknown test samples with the class identified:

First, Partial Least Squares needs to be performed for the data samples and while performing regression analysis, that PLS model will be taken into account to predict the concentrations. So the accuracy of the PLS model should be good. The data comprises of the samples to be predicted and the factors that are considered to predict is the impedance reading at the different frequencies.



Select model: PLS

Components: 5

○ Short Prediction

● Full Prediction

□ Inlier limit

□ Sample Inlier dist

☑ Identify Outliers

Outlier waming limits

Leverage Limit: 3.0

Sample Outlier Limit, Validation 2.6

Individual Value Outlier, Validation: 3.0

Total Explained Variance (%): 20

Data

Matrix: Data Matrix {31x14}

Rows: prediction {4} Cols: freq {10}

□ Include Y reference

Data

Matrix

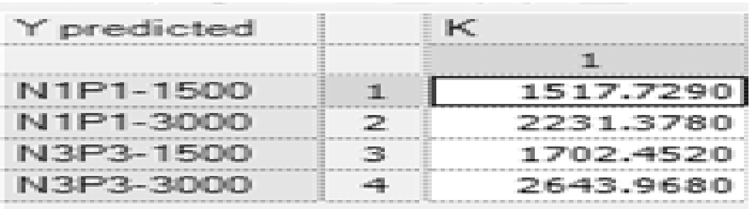
Rows Cols

6.8.1. Prediction Results:

Prerequisites for prediction of response values on new samples for which X-values are available are the following:

1. Prediction requires a regression model (MLR, PCR or PLS) which expresses the response variable(s) (Y) as a function of the X-variables.

2. The model should have been *calibrated* on samples covering the same region the new samples belong to, i.e. on similar samples (similarity being determined by the X-values).
3. The model should have been *validated* on samples covering the region the new samples belong to, preferably using test set validation, however, cross-validation can also be used.



Y predicted		K
		1
N1P1-1500	1	1517.7290
N1P1-3000	2	2231.3780
N3P3-1500	3	1702.4520
N3P3-3000	4	2643.9680

These are the predicted values for the Potassium ion for the test samples. The results are found to be approximately correct. The description of results is as follows:

SAMPLES	ACCURATE VALUES	ESTIMATED VALUES
N1P1-1500	1500	1517.7290
N1P1-3000	3000	2231.3780
N3P3-1500	1500	1702.4520
N3P3-3000	3000	2643.9680

6.9. Fill Missing Values:

It may sometimes be difficult to gather values of all the variables of interest, for all the samples included in a given study. As a consequence, some of the cells in a data table will remain empty. This may also occur if some values are lost due to human or instrumental failure, or if a recorded value appears so improbable that it must be deleted, thus creating an empty cell.

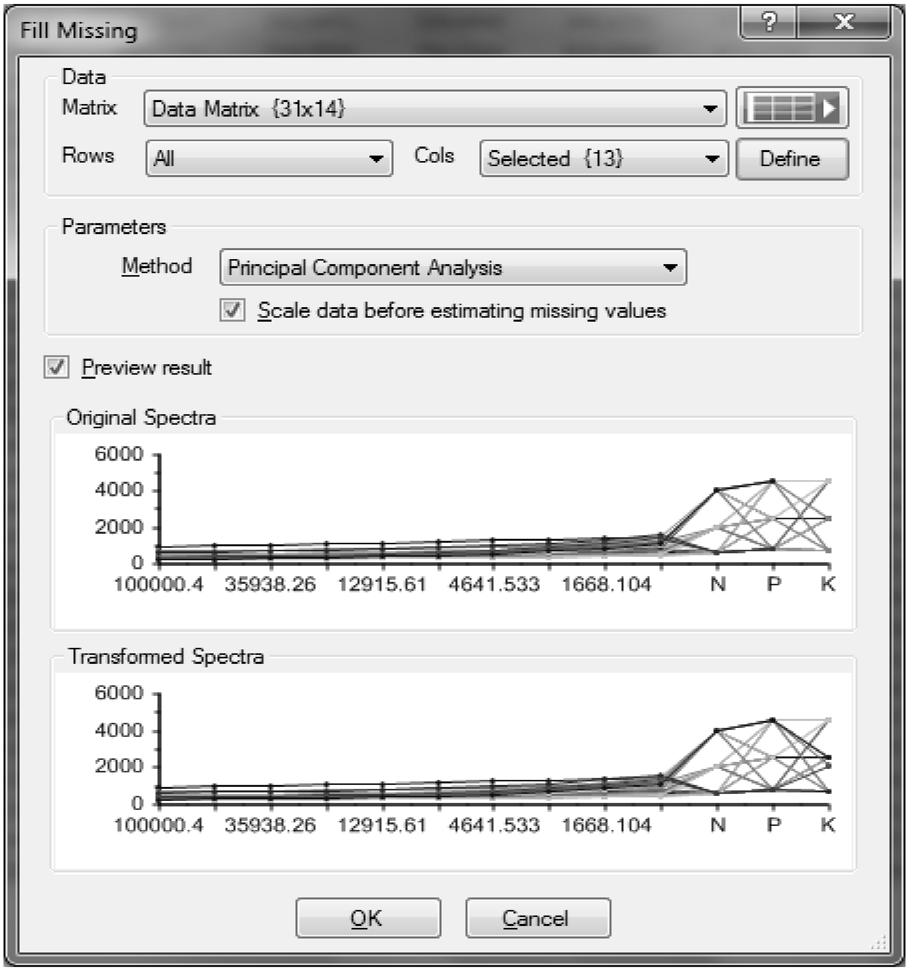
Although some of the analysis methods (PCA, PCR, PLS, MCR) available in The Unscrambler can cope with a reasonable amount of missing values, there are still multiple advantages in filling empty cells with estimated values:

- Allow all points to appear on a 2-D or 3-D scatter plot;
- Enable the use of transformations requiring that all values are non-missing, such as derivatives;

- Enable the use of analysis methods requiring that all values are non-missing, like for instance MLR or Analysis of Effects.

Steps to perform ‘Fill Missing Values’:

Step 1: Tasks → Transform → Fill Missing values



Fill Missing

Data

Matrix: Data Matrix {31x14}

Rows: All

Cols: Selected {13}

Parameters

Method: Principal Component Analysis

☑ Scale data before estimating missing values

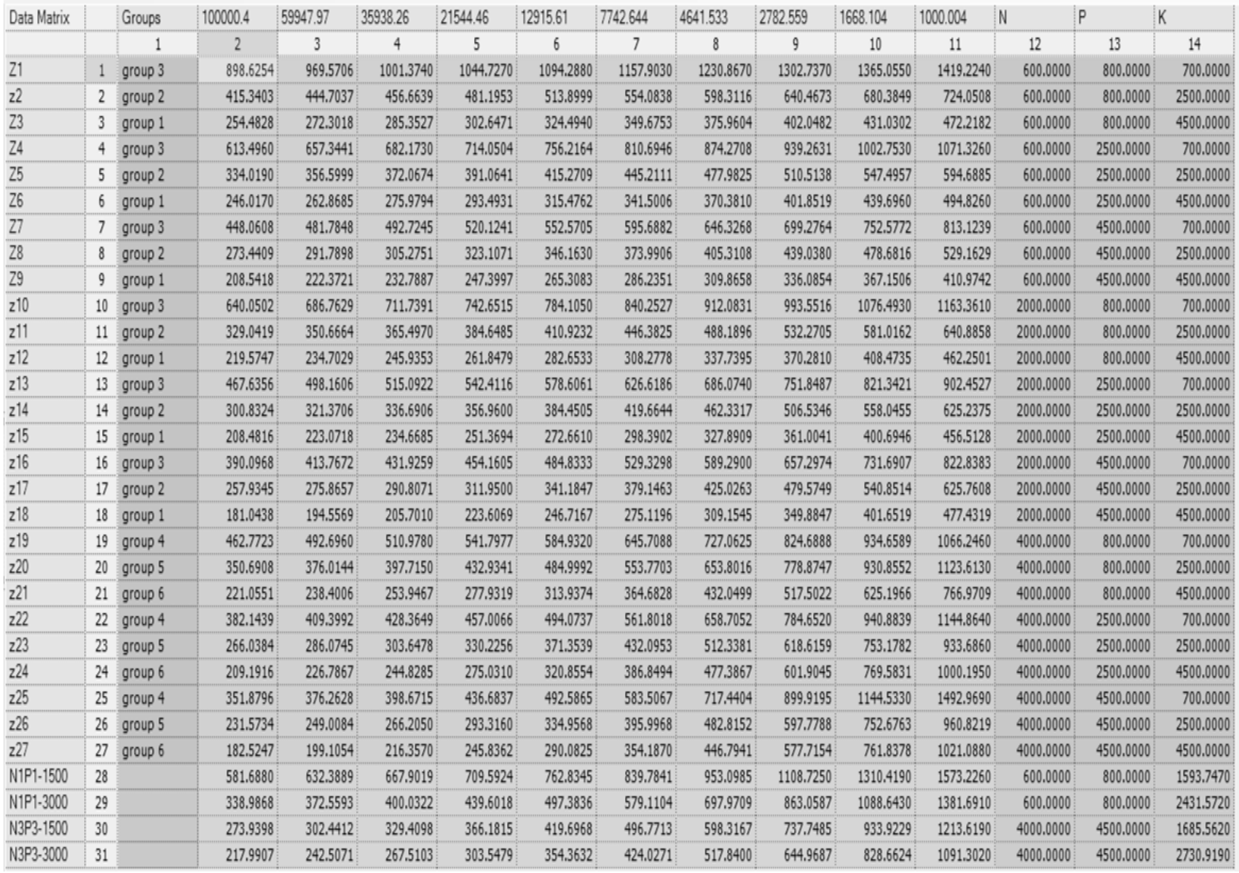
☑ Preview result

Step 2: Two methods are available for the estimation of missing values:

Principal Component Analysis performs a reconstruction of the missing values based on a PCA model of the data with an optimal number of components. This fill missing procedure is the *default* selection and the recommended method of choice for spectroscopic data.

Row Column Means Method only makes use of the *same column and row* as each cell with missing data. Use this method if the columns or rows in the data come from very different sources that do not carry information about other rows or columns. This can be the case for process data.

6.9.1. Interpreting the results:



Data Matrix		Groups	100000.4	59947.97	35938.26	21544.46	12915.61	7742.644	4641.533	2782.559	1668.104	1000.004	N	P	K
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Z1	1	group 3	898.6254	969.5706	1001.3740	1044.7270	1094.2880	1157.9030	1230.8670	1302.7370	1365.0550	1419.2240	600.0000	800.0000	700.0000
z2	2	group 2	415.3403	444.7037	456.6639	481.1953	513.8999	554.0838	598.3116	640.4673	680.3849	724.0508	600.0000	800.0000	2500.0000
Z3	3	group 1	254.4828	272.3018	285.3527	302.6471	324.4940	349.6753	375.9604	402.0482	431.0302	472.2182	600.0000	800.0000	4500.0000
Z4	4	group 3	613.4960	657.3441	682.1730	714.0504	756.2164	810.6946	874.2708	939.2631	1002.7530	1071.3260	600.0000	2500.0000	700.0000
Z5	5	group 2	334.0190	356.5999	372.0674	391.0641	415.2709	445.2111	477.9825	510.5138	547.4957	594.6885	600.0000	2500.0000	2500.0000
Z6	6	group 1	246.0170	262.8685	275.9794	293.4931	315.4762	341.5006	370.3810	401.8519	439.6960	494.8260	600.0000	2500.0000	4500.0000
Z7	7	group 3	448.0608	481.7848	492.7245	520.1241	552.5705	595.6882	646.3268	699.2764	752.5772	813.1239	600.0000	4500.0000	700.0000
Z8	8	group 2	273.4409	291.7898	305.2751	323.1071	346.1630	373.9906	405.3108	439.0380	478.6816	529.1629	600.0000	4500.0000	2500.0000
Z9	9	group 1	208.5418	222.3721	232.7887	247.3997	265.3083	286.2351	309.8658	336.0854	367.1506	410.9742	600.0000	4500.0000	4500.0000
z10	10	group 3	640.0502	686.7629	711.7391	742.6515	784.1050	840.2527	912.0831	993.5516	1076.4930	1163.3610	2000.0000	800.0000	700.0000
z11	11	group 2	329.0419	350.6664	365.4970	384.6485	410.9232	446.3825	488.1896	532.2705	581.0162	640.8858	2000.0000	800.0000	2500.0000
z12	12	group 1	219.5747	234.7029	245.9353	261.8479	282.6533	308.2778	337.7395	370.2810	408.4735	462.2501	2000.0000	800.0000	4500.0000
z13	13	group 3	467.6356	498.1606	515.0922	542.4116	578.6061	626.6186	686.0740	751.8487	821.3421	902.4527	2000.0000	2500.0000	700.0000
z14	14	group 2	300.8324	321.3706	336.6906	356.9600	384.4505	419.6644	462.3317	506.5346	558.0455	625.2375	2000.0000	2500.0000	2500.0000
z15	15	group 1	208.4816	223.0718	234.6685	251.3694	272.6610	298.3902	327.8909	361.0041	400.6946	456.5128	2000.0000	2500.0000	4500.0000
z16	16	group 3	390.0968	413.7672	431.9259	454.1605	484.8333	529.3298	589.2900	657.2974	731.6907	822.8383	2000.0000	4500.0000	700.0000
z17	17	group 2	257.9345	275.8657	290.8071	311.9500	341.1847	379.1463	425.0263	479.5749	540.8514	625.7608	2000.0000	4500.0000	2500.0000
z18	18	group 1	181.0438	194.5569	205.7010	223.6069	246.7167	275.1196	309.1545	349.8847	401.6519	477.4319	2000.0000	4500.0000	4500.0000
z19	19	group 4	462.7723	492.6960	510.9780	541.7977	584.9320	645.7088	727.0625	824.6888	934.6589	1066.2460	4000.0000	800.0000	700.0000
z20	20	group 5	350.6908	376.0144	397.7150	432.9341	484.9992	553.7703	653.8016	778.8747	930.8552	1123.6130	4000.0000	800.0000	2500.0000
z21	21	group 6	221.0551	238.4006	253.9467	277.9319	313.9374	364.6828	432.0499	517.5022	625.1966	766.9709	4000.0000	800.0000	4500.0000
z22	22	group 4	382.1439	409.3992	428.3649	457.0066	494.0737	561.8018	658.7052	784.6520	940.8839	1144.8640	4000.0000	2500.0000	700.0000
z23	23	group 5	266.0384	286.0745	303.6478	330.2256	371.3539	432.0953	512.3381	618.6159	753.1782	933.6860	4000.0000	2500.0000	2500.0000
z24	24	group 6	209.1916	226.7867	244.8285	275.0310	320.8554	386.8494	477.3867	601.9045	769.5831	1000.1950	4000.0000	2500.0000	4500.0000
z25	25	group 4	351.8796	376.2628	398.6715	436.6837	492.5865	583.5067	717.4404	899.9195	1144.5330	1492.9690	4000.0000	4500.0000	700.0000
z26	26	group 5	231.5734	249.0084	266.2050	293.3160	334.9568	395.9968	482.8152	597.7788	752.6763	960.8219	4000.0000	4500.0000	2500.0000
z27	27	group 6	182.5247	199.1054	216.3570	245.8362	290.0825	354.1870	446.7941	577.7154	761.8378	1021.0880	4000.0000	4500.0000	4500.0000
N1P1-1500	28		581.6880	632.3889	667.9019	709.5924	762.8345	839.7841	953.0985	1108.7250	1310.4190	1573.2260	600.0000	800.0000	1593.7470
N1P1-3000	29		338.9868	372.5593	400.0322	439.6018	497.3836	579.1104	697.9709	863.0587	1088.6430	1381.6910	600.0000	800.0000	2431.5720
N3P3-1500	30		273.9398	302.4412	329.4098	366.1815	419.6968	496.7713	598.3167	737.7485	933.9229	1213.6190	4000.0000	4500.0000	1685.5620
N3P3-3000	31		217.9907	242.5071	267.5103	303.5479	354.3632	424.0271	517.8400	644.9687	828.6624	1091.3020	4000.0000	4500.0000	2730.9190

These are the predicted values for the Potassium ion for the test samples through the fill missing values. The results are found to be approximately correct & in the same range as the results from the prediction by regression method. The description of results is as follows:

SAMPLES	ACCURATE VALUES	ESTIMATED VALUES
N1P1-1500	1500	1593.747
N1P1-3000	3000	2431.572
N3P3-1500	1500	1685.562
N3P3-3000	3000	2730.919

6.10. Cluster Analysis:

A valuable tool for exploratory data analysis is the use of cluster analysis to understand the natural grouping of objects. Cluster analysis is an unsupervised methodology for grouping things based on their similarities based on specified characteristics (variables). It grew out of work by biologists working on numerical taxonomy, and is a valuable visualization tool in data mining. One can perform clustering using either several agglomerative methods: K-means or K-median clustering, or hierarchical clustering with different linkage measures (single-linkage, complete-linkage, average-linkage, median-linkage, etc.). Agglomerative methods begin by treating each sample as a single cluster and begin clustering samples based on their similarity until one large cluster is formed.

6.10.1. Steps to perform:

Step 1: Define the data matrix that needs to be divided into clusters. There are 2 methods for clustering in Unscrambler:

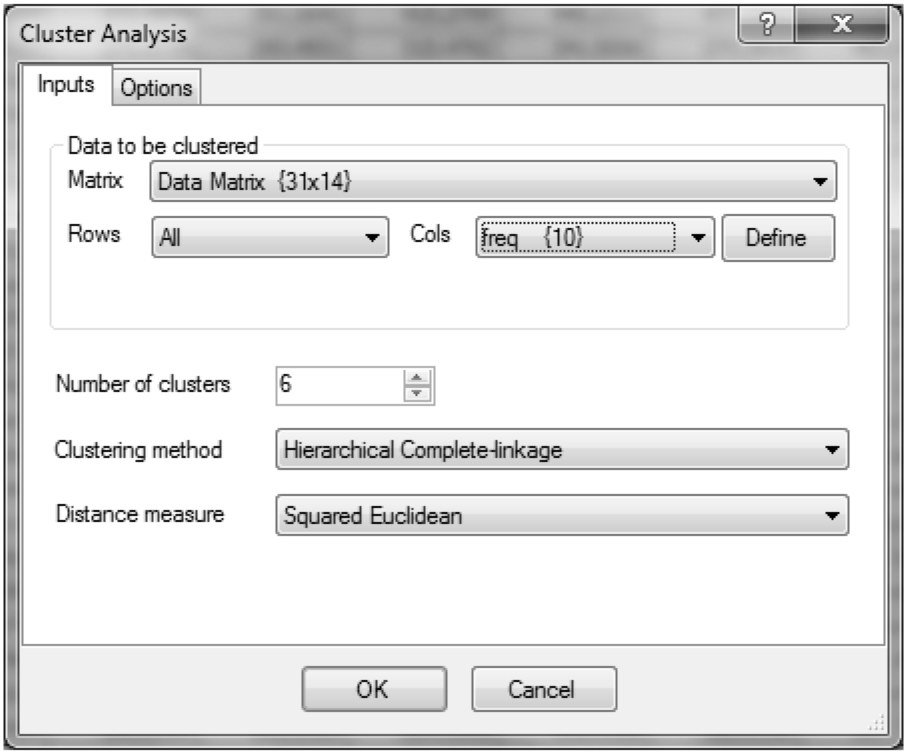
- **K Means Clustering:**

The user specifies the number of clusters in advance, and can also define cluster membership as well. The output is to give class identification for each object (sample). *K-Means methodology* is a commonly used clustering technique. The K-medians methodology is also used, and though slower than K-means, is more robust to outliers. In both cases the analysis involves starting with a collection of samples that one attempts to group them into *k* Number of clusters based on certain specific distance measurements.

- **Hierarchical Clustering (HCA):**

HCA is based on using different linkage methods to generate clusters. The user must therefore choose the linkage method, as well as the distance measure that will be used to define the clusters (separate the classes) in a data set. The distance between objects or clusters can be computed in several ways and these can have major impacts on the resulting classification.

HCA complete-linkage: This is also known as the farthest-neighbor method, and uses the greatest distance between any two samples as the basis of the clustering. Clusters from the complete-linkage method are more compact and rounded clusters.



Data to be clustered

Matrix: Data Matrix {31x14}

Rows: All

Cols: freq {10}

Number of clusters: 6

Clustering method: Hierarchical Complete-linkage

Distance measure: Squared Euclidean

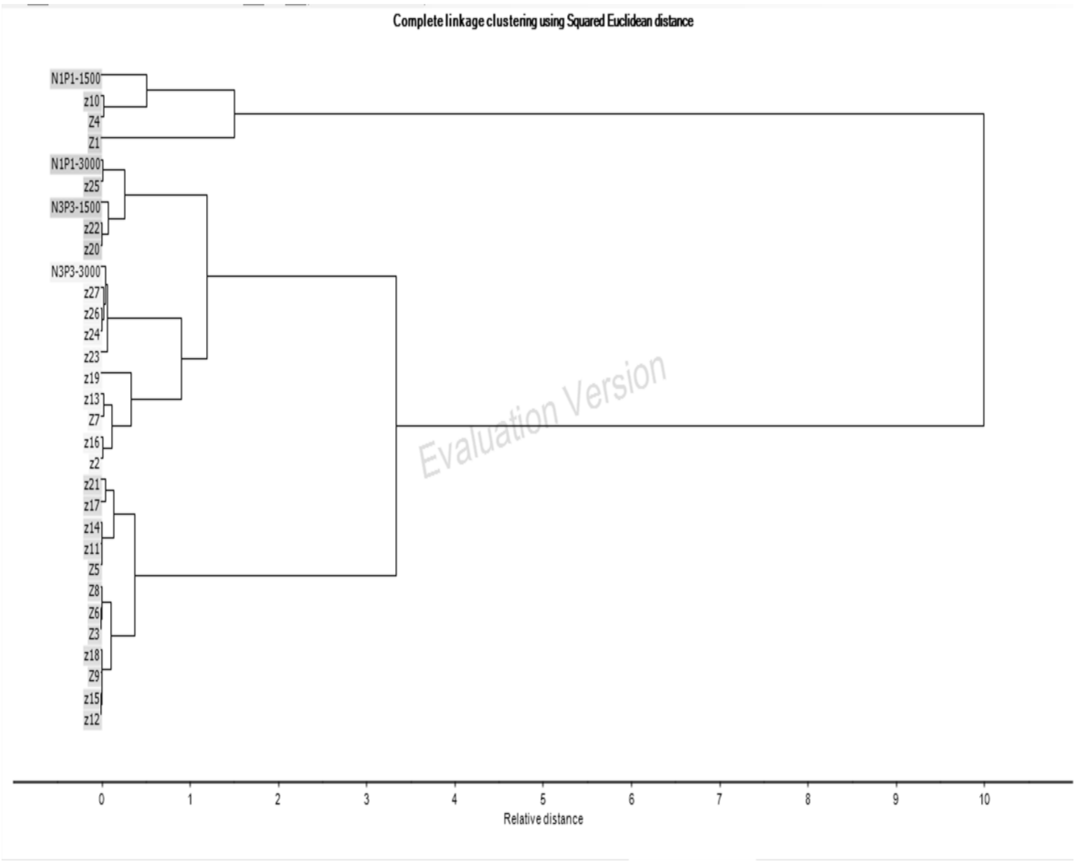
6.10.2. Interpreting the results of cluster analysis:

A dendrogram (from Greek dendron "tree", -gramma "drawing") is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Depending on the selected number of clusters, the sample names will be displayed by cluster color. In the following example three clusters were selected, hence the plot has three groups of samples shown in different colors. The clusters are separated based on the distance between clusters.

Based on the concentration of the unknown laboratory test samples, we know that the first test sample N1P1-1500 needs to be close to sample 1 or sample 2. Similarly, other unknown test samples results are as follows:

Unknown 1 (N1P1-1500):	sample 1	or	sample 2
Unknown 2 (N1P1-3000):	sample 2	or	sample 3

Unknown 3 (N3P3-1500):	sample 25	or	sample 26
Unknown 4 (N3P3-3000):	sample 26	or	sample 27



As we can see in the figure that the results are true for three unknown test samples: Unknown 1, Unknown 3 & Unknown 4. The results are true to the 75% of the accuracy. As described earlier, the result is incorrect for the Unknown 2 which may be due to the manual errors that might have happened during the experimentation times.

References

1. Earl R., Thomas G. and Blackmore. The potential role of GIS in autonomous field operations. Computers and Electronics in agriculture.25, 107-120). 2000.
2. Blackmore. The Role of Precision Farming in Sustainable Agriculture. A European Perspective. Paper presented at the International Conference on Site-Specific Management for Agriculture Systems, Minneapolis, and USA.1994.
3. Precision Farming, Anil Kumar Singh, 2008
4. Stephen W. Searcy, “Precision farming: A New Approach to Crop Management, 1997
5. Precision Farming: Dreams and Realities for Indian Agriculture, U.K. Shanwad, V.C. Patil & H. Honne Gowda, 2004
6. Statistical Pattern Recognition: A Review Ani1 K. Jain, Robert P.W. Duin, and Jianchang Mao, 2000
7. Gonzalez, R.C.Thomas, M.G .Syntactic Pattern Recognition: an Introduction ,Addison Wesley,Reading,MA,1978
8. Sergios Theodoridis, Konstantinos Koutroumbas , pattern recognition , Pattern Recognition ,Elsevier(USA)) ,1982
9. RJ Schalkoff. Pattern Recognition: Statistical, Structural and Neural Approaches. John Wiley & Sons, 1992
10. Frank T.Allen*,Jason M.kinser, H. John Gaulfield, A neural bridge from syntactic pattern recognition to statistical pattern recognition ,Neural Networks 12 ,519-526 ,1999
11. Pavilidis, T., Structural Pattern Recognition, Springer-Verlag, New York, 1977.
12. Hyeran Byun , Seong-Whan Holland Lee, Applications of Support Vector Machines for Pattern Recognition: A Survey, SVM 2002, LNCS 2388, pp. 213-236 , 2002
13. Pattern Recognition: An overview, Jie Liu, Jigui Sun, Shengsheng Wang, 2006
14. Classification Techniques in Pattern Recognition, Lihong Zheng and Xiangjian He, 2007
15. Statistical Approach to Clustering in Pattern Recognition, Yujing Zeng and Janusz Starzyk, 2008
16. Discrimination of teas based on total luminescence spectroscopy and pattern recognition, L. Nitin Seetohul, Meez Islam, William T O’Hare, Zulfiqur Ali, 2006

17. A novel iTongue for Indian black tea discrimination, Amol P. Bhondekar, Mopsy Dhiman, Anupma Sharma, Arindam Bhaktaa, Abhijit Gangulib, S.S. Baric, Renu Vigc, Pawan Kapur, Madan L. Singla, 2010
18. Soil pattern recognition with fuzzy c-means: application to classification and soil-landform interrelationship, Odeh, McBratney, Chittleborough, D.J., 1992.
19. Performance Evaluation of a Novel iTongue for Indian Black Tea Discrimination, Amol P. Bhondekar, Renu Vig, Ashu Gulati, Madan L. Singla, and Pawan Kapur, 2011
20. J. Kaiser, Science, 2001, 294, 1268–1269.
21. P. A. Vadas, P. J. A. Kleinman and A. N. Sharpley, J. Environ. Qual., 2004, 33, 749–756.
22. T. Page, P. M. Haygarth, K. J. Beven, A. Joynes, T. Butler, C. Keeler, J. Freer, P. N. Owens and G. A. Wood, J. Environ. Qual., 2005, 34, 2263–2277.
23. M. L. Ruffo, G. A. Bollero, R. G. Hoeft and D. G. Bullock, Agron. J., 2005, 97, 1485–1492.
24. K. A. Sudduth, J. W. Hummel and S. J. Birrell, in The State of Site-Specific Management for Agriculture, ed. F. J. Pierce and E. J. Sadler,ASA-CSSA-SSSA, Madison, WI, 1997, pp.183–210.
25. N. R. Kitchen, K. A. Sudduth, D. B. Myers, R. E. Massey, E. J. Sadler, R. N. Lerch, J. W. Hummel and H. L. Palm, J. Soil Water Conserve., 2005, 60, 421–430.
26. J. S. Schepers and M. R. Schlemmer, Proceedings of the 1st International Conference on Geospatial Information in Agriculture and Forestry, Ann Arbor, 1998.
27. N. C. Wollenhaupt, D. J. Mulla and C. A. G. Crawford, in The State of Site-Specific Management for Agriculture, ed. F. J. Pierce and E. J. Sadler, ASA-CSSA-SSSA, Madison, WI, 1997, pp.19–53.
28. V. I. Adamchuk, J. W. Hummel, M. T. Morgan and S. K. Upadhyaya, Compute. Electron. Agric., 2004, 44, 71–91.
29. S. J. Birrell and J. W. Hummel, Trans. ASAE, 2000, 43, 197–206.
30. S. D. Moss, J. Janata and C. C. Johnson, Anal. Chem., 1975, 47, 2238–2242.
31. M. Knoll, K. Cammann, C. Dumschat, M. Borchardt and G. Hogg, Sens. Actuators, B, 1994, 20, 1–5.
32. Soil macronutrient sensing for precision agriculture, Hak-Jin Kim,*a Kenneth A. Sudduthb and John W. Hummelb, 2009

33. On-the-go soil sensors for precision agriculture, V.I. Adamchuka,∗, J.W. Hummelb, M.T. Morgan c, S.K. Upadhyaya, 2004

34. Morgan, M.T., Ess, D.R., 1997. The Precision-Farming Guide for Agriculturists. An agriculture primer. John Deere Publishing, Moline, IL.

35. Sudduth, K.A., Hummel, J.W., Birrell, S.J., 1997. Sensors for site-specific management. In: Pierce, F.T., Sadler, E.J. (Eds.), The State of Site-Specific Management for Agriculture, ASA- CSSA-SSSA, Madison, Wisconsin, Chapter 10, pp. 183–210.

36. Baumgardner, M.F., Silva, L.F., Beihl, L.L., Stoner, E.R., 1985. Reflectance properties of soils. Advances in Agronomy 38, 1–44.

37. Birrell, S.J., Hummel, J.W., 1997. Multi-sensor ISFET system for soil analysis. In: Stafford, J.V. (Ed.), Proceedings of the First European Conference on Precision Agriculture. BIOS Scientific Publishers Ltd., Oxford, UK, pp. 459–468.

38. Adsett, J.F., Thottan, J.A., Sibley, K.J., 1999. Development of an automated on-the-go soil nitrate monitoring system. Applied Engineering in Agriculture 15 (4), 351–356.

39. L. Yongguo *et al., Finding the Optimal Number of Clusters Using Genetic Algorithms*, pp. 1325–1330, 2008.

40. D. L. Davies and D. W. Bouldin, “A cluster separation measure,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, 1979.

41. Data Clustering and Its Applications Raza Ali , Usman Ghani, Aasim Saeed, 2008

42. Title: Practical Handbook of Disturbed Land Revegetation Date: 1994 Author: Munshower, F.F.

43. Title: The Nature and Properties of Soils Date: 1999 Author: Brady, N.C., and R.R. Weil

44. Title: Methods of Soil Analysis Part 2 - Chemical and Microbiological Properties Date: 1982 Author: Page, A.L., R.H. Miller, and D.R. Keeney

45. Jie Wu Æ Yuxing Ben Æ Hsueh-Chia Chang Particle detection by electrical impedance spectroscopy with asymmetric-polarization AC electro osmotic trapping, 2005

46. Impedance Spectroscopy J. ROSS Macdonald, 1991

47. A. Riul, H.C. de Sousa, R.R. Malmegrim, D.S. dos Santos, A.C.P.L.F. Carvalho, F.J.Fonseca, O.N. liveira, L.H.C. Mattoso, Wine classification by taste sensors made from ultra-thin films and using neural networks, Sensors and Actuators B: Chemical 98 (2004) 77–82.